

Concentration inequalities for Markov chains by Marton couplings

Daniel Paulin

Daniel Paulin

*Department of Mathematics, National University of Singapore
10 Lower Kent Ridge Road, Singapore 119076, Republic of Singapore.
e-mail: paulindani@gmail.com*

Abstract: We use a coupling construction of Katalin Marton to prove Hoeffding, Bernstein, Mcdiarmid, Talagrand, and self-bounding type concentration inequalities for Markov chains with countable state space, in discrete, and continuous time. The constants in the exponents are roughly the “mixing time of the Markov chain” times worse than in the independent case.

AMS 2000 subject classifications: Primary 60E15, 60J10, 60J27, 28A35; secondary 05C81., 68Q87.

Keywords and phrases: Concentration inequality, coupling, Markov chain, mixing time, spectral gap.

Contents

| | | |
|-----|---|----|
| 1 | Introduction | 1 |
| 1.1 | Main definitions | 2 |
| 1.2 | Additional definitions | 6 |
| 2 | Results | 7 |
| 2.1 | Results by Marton couplings | 7 |
| 2.2 | Results by spectral methods | 14 |
| 2.3 | Results for continuous time chains | 15 |
| 3 | Applications | 16 |
| 3.1 | Coin tossing | 16 |
| 3.2 | Error analysis for MCMC | 18 |
| 3.3 | m -dependence | 19 |
| 3.4 | Independent permutations and Bernoulli variables with fixed sum | 19 |
| 3.5 | Hidden Markov chains | 20 |
| 3.6 | Random walks on weighted graphs | 21 |
| 4 | Open problems | 23 |
| 5 | Proofs | 26 |
| | Acknowledgements | 53 |
| | References | 53 |

1. Introduction

For readers not familiar with concentration inequalities, we recommend [Ledoux \(2001\)](#), [Mitzenmacher and Upfal \(2005\)](#) and [Dubhashi and Panconesi \(2009\)](#).

The transportation inequality cost method to prove measure concentration was initiated by Katalin Marton. [Marton \(1986\)](#) proves the blow-up lemma (a weaker version of measure concentration in Hamming distance).

[Marton \(1996a\)](#) proves measure concentration in Hamming distance for countable state Markov chains. For a homogeneous Markov chain with state space Ω , and transition probabilities $P_{i,j}$, let us denote

$$a := \max_{i,j \in \Omega} d_{TV}(P_{i,\cdot}, P_{j,\cdot}), \quad (1.1)$$

then Proposition 1 of [Marton \(1996a\)](#) proves that measure concentration holds with constants $1/(1-a)^2$ times worse than in the independent case.

[Marton \(1996b\)](#), [Marton \(1997\)](#) extends this result, and proves Talagrand's inequality for Markov chains, with constants $1/(1-a)^2$ times worse than in the independent case.

The Markov chain setting was further generalized to a class of random processes, for Hamming distance, in [Marton \(1998a\)](#).

Talagrand's convex distance inequality, for a larger class of random processes, was independently proven in [Marton \(1998b\)](#) and [Samson \(2000\)](#). The latter also proves a weak version of Talagrand's suprema of empirical processes inequality.

In [Marton \(2003\)](#), these results are further extended to prove concentration inequalities for a larger class of functions.

[Chazottes et al. \(2007\)](#) (using an elementary martingale-type argument) and [Kontorovich \(2007\)](#) (using martingales and linear algebraic inequalities) prove concentration inequalities, in Hamming distance, for a class of mixing coefficients, similar to those of [Samson \(2000\)](#). For homogeneous countable state Markov chains, their results are the same as Proposition 1 of [Marton \(1996a\)](#).

[Lezaud \(1998a\)](#) proves Bernstein - type concentration inequalities for finite state Markov chains, for empirical means $n^{-1} \sum_{i=1}^n f(X_i)$ (and $\frac{1}{t} \int_{s=0}^t f(X_s) ds$ in the continuous case). For reversible Markov chains, the constants in the exponents depend on the spectral gap of the chain. [Lezaud \(2001\)](#) generalizes this to Markov processes with general state space, and proves a Berry-Esseen bound for the empirical mean.

The purpose of this paper is improve these results, and show that concentration inequalities for Markov chains are in fact governed by the mixing time of the chain. This work grew out of the author's attempt to solve the "Spectral transportation cost inequality" conjecture stated in Section 6.4. of [Kontorovich \(2007\)](#).

1.1. Main definitions

In the following, we will consider dependent random variables $X = (X_1, \dots, X_N)$ taking values in some set

$$\Lambda := \Lambda_1 \times \dots \times \Lambda_N,$$

and let P also denote the law of X , i.e. $X \sim P$. Let $Y := (Y_1, \dots, Y_N)$ be a random vector taking values in Λ , and suppose $Y \sim Q$. Denote $[N] := \{1, \dots, N\}$.

Assumption 1.1. *For notational convenience, we will suppose that Λ is discrete. The continuous case can be treated similarly, as it is done in [Samson \(2000\)](#).*

We will denote a coupling of P and Q by

$$\pi[X \sim P, Y \sim Q]. \quad (1.2)$$

An example: let $\pi[X \sim P, Y \sim Q]$ be the maximal coupling of P and Q (see [Lindvall \(1992\)](#)), then

$$\pi[X \neq Y] = d_{TV}(P, Q).$$

We will need to refer to subsets of our vectors, let

$$X_{\leq k} := (X_1, \dots, X_k), \quad X_{\geq k}^n := (X_k, \dots, X_n). \quad (1.3)$$

The laws of these “subvectors” will be denoted by $P_{\leq k}$, and $P_{\geq k}^n$, respectively.

The following is the most important definition of this paper. It has appeared in [Marton \(2003\)](#).

Definition 1 (Marton coupling). *Let $\mathcal{X} := (\mathcal{X}_1, \dots, \mathcal{X}_N)$ be a vector of random variables taking values in $\Omega = \Omega_1 \times \dots \times \Omega_N$, with law \mathcal{P} . We define a Marton coupling for \mathcal{X} as a set of couplings*

$$\mathcal{M} := \left\{ \mathcal{M}^i \left[\mathcal{X}_{\geq i+1}^N \sim \mathcal{P}_{\geq i+1}^N(\cdot | x_{\leq i}), \mathcal{X}'_{\geq i+1} \sim \mathcal{P}_{\geq i+1}^N(\cdot | x_{\leq i-1}, x'_i) \right] \right\}_{i \leq N},$$

i.e. for each $i \leq N$ and $(x_{\leq i}, x'_i)$, $\mathcal{M}^i := \mathcal{M}^i(\cdot | x_{\leq i}, x'_i)$ is a coupling between $\mathcal{X}_{\geq i+1}^N \sim \mathcal{P}_{\geq i+1}^N(\cdot | x_{\leq i})$ and $\mathcal{X}'_{\geq i+1} \sim \mathcal{P}_{\geq i+1}^N(\cdot | x_{\leq i-1}, x'_i)$, satisfying the following condition:

$$\text{for every } (x_{\leq i}, x'_i) \text{ with } x_i = x'_i, \mathcal{M}^i[\mathcal{X}_{\geq i+1}^N = \mathcal{X}'_{\geq i+1} | x_{\leq i}, x'_i] = 1. \quad (1.4)$$

We define the mixing matrix of \mathcal{M} , $\Gamma := (\Gamma_{i,j})_{i,j \leq N}$ as an upper diagonal matrix with

$$\Gamma_{i,i} := 1 \text{ for } i \leq N, \text{ and } \Gamma_{i,j} := \sup_{x_{\leq i}, x'_i} \mathcal{M}^i[\mathcal{X}_j \neq \mathcal{X}'_j | x_{\leq i}, x'_i] \text{ for } 1 \leq i < j \leq N.$$

Remark 1.1. [Samson \(2000\)](#), [Chazottes et al. \(2007\)](#), [Chazottes and Redig \(2009\)](#), [Kontorovich \(2007\)](#) use a similar construction, but assume that \mathcal{M}^i is the maximal coupling between

$$P_{\geq i+1}^N(\cdot | x_{\leq i}) \text{ and } P_{\geq i+1}^N(\cdot | x_{\leq i-1}, x'_i)$$

(the coupling that “achieves” the total variation distance, see [Definition 8](#), or [Lindvall \(1992\)](#)). We will use the extra freedom provided by [Definition 1](#) for our theorems in this paper.

For homogeneous Markov chains, and M^i defined as the maximal coupling, we get

$$\Gamma = (\Gamma_{i,j})_{i,j \leq n} \leq \begin{pmatrix} 1 & a & a^2 & a^3 & \dots \\ 0 & 1 & a & a^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \quad (1.5)$$

which gives $\|\Gamma\| < \frac{1}{1-a}$ (a is defined as in [\(1.1\)](#)).

Caveat lector: although it is true that

$$\Gamma_{i,j} \geq \max_{x_{\leq i}, x'_i} d_{TV}(P_j(\cdot | x_{\leq i}), P_j(\cdot | x_{\leq i-1}, x'_i)),$$

the equality does not holds in general.

We will use the definition of partition of a set:

Definition 2 (Partition). A partition of a set S is the division of S into disjoint non-empty subsets that together cover S . Analogously, we say that $\hat{X} = (\hat{X}_1, \dots, \hat{X}_n)$ is a partition of a set of random variables $X := (X_1, \dots, X_N)$ if $(\hat{X}_i)_{1 \leq i \leq n}$ are disjoint, and together cover X . For a partition \hat{X} of X , we denote the number of elements \hat{X}_i by $s(\hat{X}_i)$ (size of \hat{X}_i), and call $s(\hat{X}) := \max_{1 \leq i \leq n} s(\hat{X}_i)$ the size of the partition.

Finally, we denote the set of indices of the elements of \hat{X}_i by $\mathcal{I}(\hat{X}_i)$, i.e. $X_j \in \hat{X}_i$ if and only if $j \in \mathcal{I}(\hat{X}_i)$. For a set of indices $S \subset [N]$, let $X_S := \{X_j : j \in S\}$. In particular, $\hat{X}_i = X_{\mathcal{I}(\hat{X}_i)}$.

The state space of X will be denoted by $\Lambda := \Lambda_1 \times \dots \times \Lambda_N$ and the state space of \hat{X} is denoted by $\hat{\Lambda} := \hat{\Lambda}_1 \times \dots \times \hat{\Lambda}_n$, with $\hat{\Lambda}_i = \Lambda_{\mathcal{I}(\hat{X}_i)}$.

In Section 4.5 and 4.6 of [Levin, Peres and Wilmer \(2009\)](#), the mixing time of a time homogeneous chain is defined the following way:

Definition 3. Let X_1, X_2, X_3, \dots be a countable state, time homogeneous Markov chain with transition matrix P , state space Ω , and stationary distribution π .

Let us denote

$$d(t) := \sup_{x \in \Omega} d_{TV}(P^t(x, \cdot), \pi),$$

$$t_{mix}(\epsilon) := \min\{t : d(t) \leq \epsilon\}$$

and

$$t_{mix} := t_{mix}(1/4).$$

We will use the following alternative definition, which also works for time inhomogeneous Markov chains:

Definition 4. Let X_1, \dots, X_N be a countable state Markov chain with state space $\Omega_1 \times \dots \times \Omega_N$ (i.e. $X_i \in \Omega_i$). Let us denote the minimal t such that $P_{i+t}(\cdot | X_i = x)$ and $P_{i+t}(\cdot | X_i = y)$ are less than ϵ away in total variational distance for every $1 \leq i \leq N - t$ and $x, y \in \Omega_i$ by $\tau(\epsilon)$, i.e. for $0 < \epsilon < 1$,

$$\tau(\epsilon) := \min \left\{ t \in \mathbb{N} : \max_{1 \leq i \leq N-t} \left(\sup_{x, y \in \Omega_i} d_{TV}(P_{i+t}(\cdot | X_i = x), P_{i+t}(\cdot | X_i = y)) \right) \leq \epsilon \right\}.$$

Remark 1.2. One can easily see that in the case of homogeneous Markov chains, by triangle inequality, one has

$$\tau(2\epsilon) \leq t_{mix}(\epsilon) \leq \tau(\epsilon).$$

In the following, based on Section 20 of [Levin, Peres and Wilmer \(2009\)](#), we briefly review some definitions about finite (or countable) state Markov chains with continuous time.

Let $(\Phi_k)_{k=0}^\infty$ be a time homogeneous Markov chain with transition matrix P , and countable state space Ω . Let $(T_i)_{i=1}^\infty \sim \exp(1)$ be i.i.d. exponentially distributed random variables independent of $(\Phi_k)_{k=0}^\infty$. Let $S_k = \sum_{i=1}^k T_i$ for $k \geq 1$, and define

$$X_t := \Phi_k \text{ for } S_k \leq t \leq S_{k+1}.$$

The *heat kernel* H_t is defined as

$$H_t(x, y) := \mathbb{P}(X_t = y | X_0 = x), \quad (1.6)$$

the one can show, that in matrix form, for finite state space Ω , we can express H_t with the matrix exponential:

$$H_t = \exp(t(P - \mathbf{I})), \quad (1.7)$$

here \mathbf{I} denotes the $|\Omega| \times |\Omega|$ identity matrix.

Theorem 20.1 of [Levin, Peres and Wilmer \(2009\)](#) proves that for irreducible P , there exists a stationary distribution π for X_t . We define the mixing time of continuous time homogeneous chains as in (20.7) of [Levin, Peres and Wilmer \(2009\)](#):

Definition 5.

$$t_{mix}^{cont}(\epsilon) := \inf \left\{ t \geq 0 : \sup_{x \in \Omega} d_{TV}(H_t(x, \cdot), \pi) \leq \epsilon \right\}, \quad (1.8)$$

and denote $t_{mix}^{cont} := t_{mix}^{cont}(1/4)$.

For time inhomogeneous, continuous time chains with countable state space Ω , we denote

$$H_{t_1, t_2}(x, y) := \mathbb{P}(X_{t_2} = y | X_{t_1} = x), \quad (1.9)$$

and define the mixing time analogously to Definition 4:

Definition 6.

$$\tau^{cont}(\epsilon) := \min \left\{ t \geq 0 : \sup_{s \geq 0} \sup_{x, y \in \Omega} d_{TV}[H_{s, s+t}(x, \cdot), H_{s, s+t}(y, \cdot)] \right\}, \quad (1.10)$$

and denote $\tau^{cont} := \tau^{cont}(1/4)$.

For finite state, reversible, aperiodic, irreducible chains, in discrete time, write the eigenvalues of the transition matrix P as

$$1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_{|\Omega|} \geq -1.$$

Denote

$$\lambda_* := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P, \lambda \neq 1\}, \gamma_* := 1 - \lambda_*, \gamma := 1 - \lambda_2.$$

We call γ_* the *absolute spectral gap*, and γ the *spectral gap*. Obviously, $\gamma \geq \gamma_*$. The relation between the mixing time and the spectral gap is given by the following proposition:

Proposition 1.1. *For reversible, irreducible, aperiodic chains in discrete time with finite state space Ω , we have*

$$t_{mix}(\epsilon) \geq \left(\frac{1}{\gamma_*} - 1 \right) \log \left(\frac{1}{2\epsilon} \right) \geq \left(\frac{1}{\gamma} - 1 \right) \log \left(\frac{1}{2\epsilon} \right), \quad (1.11)$$

$$t_{mix}(\epsilon) \leq \left\lceil \frac{1}{\gamma_*} \log \left(\frac{\sqrt{|\Omega|}}{\epsilon} \right) \right\rceil. \quad (1.12)$$

For continuous time, time homogeneous chains with reversible, irreducible P , and finite state space Ω , similar results hold:

$$t_{\text{mix}}^{\text{cont}}(\epsilon) \geq \left(\frac{1}{\gamma_*} - 1 \right) \log \left(\frac{1}{2\epsilon} \right) \geq \left(\frac{1}{\gamma} - 1 \right) \log \left(\frac{1}{2\epsilon} \right), \quad (1.13)$$

$$t_{\text{mix}}^{\text{cont}}(\epsilon) \leq \left\lceil \frac{1}{\gamma_*} \log \left(\frac{\sqrt{|\Omega|}}{\epsilon} \right) \right\rceil. \quad (1.14)$$

Proof. (1.11) follows by Theorem 12.4 of [Levin, Peres and Wilmer \(2009\)](#). (1.12) is proven, for example, in [Chawla \(2010\)](#). (1.13) and (1.14) are left to the reader as exercise. \square

1.2. Additional definitions

In this section we introduce some additional notations, that will be used in the statement of some of our theorems. For a (not necessarily time homogenous) Markov chain X_1, \dots, X_N , denote

$$\tau_{\min} := \inf_{0 \leq \epsilon < 1} \tau(\epsilon)/(1 - \epsilon)^2 \quad (1.15)$$

$$\tau'_{\min} := \inf_{0 \leq \epsilon < 1} \tau(\epsilon)/(1 - \sqrt{\epsilon})^2. \quad (1.16)$$

For time homogeneous chains, for some integer $t_0 \geq 0$, denote

$$\eta_{\min}(t_0) := \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor} \cdot \frac{t_{\text{mix}}(\epsilon)}{1 - \epsilon}. \quad (1.17)$$

In the time continuous case, we define $\tau_{\min}^{\text{cont}}$, $\tau'_{\min}^{\text{cont}}$, and $\eta_{\min}^{\text{cont}}(t_0)$ analogously. The following proposition gives some estimates on these quantities:

Proposition 1.2. *For time homogeneous chains, the following inequalities hold:*

$$\tau_{\min} \leq \inf_{0 \leq \epsilon < 1} t_{\text{mix}}(\epsilon/2)/(1 - \epsilon)^2 \leq \frac{128}{49} t_{\text{mix}} \leq 2.62 t_{\text{mix}}, \quad (1.18)$$

$$\tau'_{\min} \leq \inf_{0 \leq \epsilon < 1} t_{\text{mix}}(\epsilon/2)/(1 - \sqrt{\epsilon})^2 \leq 4.43 t_{\text{mix}}, \quad (1.19)$$

$$\eta_{\min}(t_0) \leq 4^{-\lfloor \frac{t_0}{t_{\text{mix}}} \rfloor} \cdot \frac{4}{3} t_{\text{mix}}. \quad (1.20)$$

The same inequalities hold in the continuous case, with τ_{\min} replaced by $\tau_{\min}^{\text{cont}}$, η_{\min} replaced by $\eta_{\min}^{\text{cont}}$, and t_{mix} replaced by $t_{\text{mix}}^{\text{cont}}$.

Remark 1.3. *In many cases, the Markov chain exhibits a cutoff, i.e. the total variation distance decreases very rapidly in a small interval, see Figure 1 of [Lubetzky and Sly \(2009\)](#). If this happens, then $\tau_{\min} \approx \tau'_{\min} \approx t_{\text{mix}}$, and $\eta_{\min}(t_0)$ decreases very quickly for $t_0 > t_{\text{mix}}$.*

Proof. The first inequality in (1.18) follows by triangle inequality, the second one by taking $\epsilon = 1/8$, and noticing that $t_{\text{mix}}(1/16) \leq 2t_{\text{mix}}$. The other inequalities are similar. \square

2. Results

2.1. Results by Marton couplings

For our first result, we need to define a distance:

Definition. For $C \in \mathbb{R}_+^{\mathcal{N}}$, $x, y \in \Omega_1 \times \dots \times \Omega_N$, we say that the C weighted Hamming distance of x and y is

$$d_C(x, y) := \sum_{i=1}^{\mathcal{N}} C_i \mathbb{1}[x_i \neq y_i], \quad (2.1)$$

and the C weighted Hamming distance of two measures, P and Q on Ω is

$$d_C(P, Q) := \inf_{\pi(X \sim P, Y \sim Q)} \sum_{i=1}^{\mathcal{N}} C_i \pi[X_i \neq Y_i]. \quad (2.2)$$

The relative entropy of P and Q will be denoted by

$$D(Q||P) := \sum_{x \in \Lambda} P(x) \log \left(\frac{P(x)}{Q(x)} \right). \quad (2.3)$$

Theorem 2.1. Let $X = (X_1, \dots, X_N)$ be a sequence of random variables, $X \in \Lambda$, $X \sim P$. Let $\hat{X} = (\hat{X}_1, \dots, \hat{X}_n)$ be a partition of this sequence, $\hat{X} \in \hat{\Lambda}$, $\hat{X} \sim \hat{P}$. Suppose that we have a Marton coupling for \hat{X} with mixing matrix Γ . Then for any distribution Q on Λ , any $c \in \mathbb{R}_+^N$, we have

$$d_c(Q, P) \leq \|\Gamma \cdot C(c)\| \sqrt{\frac{1}{2} D(Q||P)}, \quad (2.4)$$

with $C(c) \in \mathbb{R}_+^n$ defined as

$$C_i(c) := \sum_{j \in \mathcal{I}(\hat{X}_i)} c_j \text{ for } i \leq n. \quad (2.5)$$

Corollary 2.1 (McDiarmid's bounded differences inequality).

Let $X, \hat{X}, \mathcal{M}, \Gamma$ and $C(c)$ as in Theorem 2.1. Let $f : \Lambda \rightarrow \mathbb{R}$ be a d_c Lipschitz function (i.e. $f(x) - f(y) \leq d_c(x, y)$) for some $c \in \mathbb{R}_+^N$, then for any $\lambda \in \mathbb{R}$,

$$\log \mathbb{E} \left(e^{\lambda(f(X) - \mathbb{E}f(X))} \right) \leq \frac{\lambda^2 \cdot \|\Gamma \cdot C(c)\|^2}{8} \leq \frac{\lambda^2 \cdot \|\Gamma\|^2 \|c\|^2 s(X)}{8}. \quad (2.6)$$

In particular, this means that

$$\mathbb{P}(f(X) \geq \mathbb{E}f(X) + t), \mathbb{P}(f(X) \leq \mathbb{E}f(X) - t) \leq \exp \left(\frac{-2t^2}{\|\Gamma \cdot C(c)\|^2} \right). \quad (2.7)$$

Corollary 2.2 (McDiarmid's inequality for Markov chains). Let $X := (X_1, \dots, X_N)$ be a (not necessarily time homogeneous) countable state Markov chain, taking values in state space $\Lambda = \Lambda_1 \times \dots \times \Lambda_N$, with mixing time $\tau(\epsilon)$.

Let $f : \Lambda \rightarrow \mathbb{R}$ be a d_c Lipschitz function for some $c \in \mathbb{R}_+^N$, then for any $\lambda \in \mathbb{R}$,

$$\log \mathbb{E} \left(e^{\lambda(f(X) - \mathbb{E}f(X))} \right) \leq \frac{\lambda^2 \cdot \|c\|^2 \cdot \tau_{\min}}{2}. \quad (2.8)$$

This implies that

$$\mathbb{P}(f(X) \geq \mathbb{E}f(X) + t), \mathbb{P}(f(X) \leq \mathbb{E}f(X) - t) \leq \exp \left(\frac{-t^2}{2\|c\|^2 \tau_{\min}} \right). \quad (2.9)$$

Corollary 2.3 (Hoeffding inequality). *Let X, \hat{X}, \mathcal{M} , and Γ as in Theorem 2.1. Suppose that $f_i : \Lambda_i \rightarrow [a_i, b_i]$, $i \leq N$. Let $c_i := b_i - a_i$, and define $C(c)$ as (2.5). Define $S := \sum_{i=1}^N f_i(X_i)$, then*

$$\mathbb{P}(S \geq \mathbb{E}S + t), \mathbb{P}(S \leq \mathbb{E}S - t) \leq \exp \left(\frac{-2t^2}{\|\Gamma \cdot C(c)\|^2} \right). \quad (2.10)$$

Corollary 2.4 (Hoeffding inequality for Markov chains). *First let $X = (X_1, \dots, X_N)$ be a (not necessarily time homogeneous) Markov chain taking values in some countable space $\Lambda := \Lambda_1 \times \dots \times \Lambda_N$. Suppose that $f_i : \Lambda_i \rightarrow [a_i, b_i]$, $i \leq N$. Define $S := \sum_{i=1}^N f_i(X_i)$, then*

$$\mathbb{P}(S \geq \mathbb{E}S + t), \mathbb{P}(S \leq \mathbb{E}S - t) \leq \exp \left(\frac{-2t^2}{\tau_{\min} \sum_{i=1}^N (a_i - b_i)^2} \right).$$

Now suppose, in addition, that X is time homogeneous, and $\Lambda_1 = \dots = \Lambda_n = \Omega$. Suppose that $f : \Omega \rightarrow [a, b]$. Let $t_0 \geq 0$ (“burn-in time”), and denote $Z := \frac{\sum_{i=t_0+1}^N f(X_i)}{N-t_0}$.

Then for every $t \geq 0$,

$$\begin{aligned} & \mathbb{P} \left(Z \geq \mathbb{E}_\pi(f) + \frac{(b-a)\eta_{\min}(t_0)}{N-t_0} + t \right), \mathbb{P} \left(Z \leq \mathbb{E}_\pi(f) - \frac{(b-a)\eta_{\min}(t_0)}{N-t_0} - t \right) \\ & \leq \exp \left(\frac{-2(N-t_0)t^2}{(b-a)^2 \cdot \tau_{\min}} \right). \end{aligned} \quad (2.11)$$

For our second theorem, we will need to define the d_2 distance of two measures on Λ (as in Samson (2000), and Marton (2003)):

Definition. Let P, Q be two measures on Λ , then their d_2 distance is

$$d_2(P, Q) := \inf_{\pi(X \sim P, Y \sim Q)} \left[\sum_{y \in \Lambda} \sum_{i=1}^N \pi[X_i \neq y_i | Y_i = y_i]^2 \cdot Q(y) \right]^{1/2} \quad (2.12)$$

$$= \inf_{\pi(X \sim P, Y \sim Q)} \sup_{\alpha: E_Q(\sum_i \alpha_i^2(Y)) \leq 1} \mathbb{E}_\pi \left(\sum_{i=1}^N \alpha_i(Y) \mathbb{1}[X_i \neq Y_i] \right), \quad (2.13)$$

where $\alpha : \Lambda \rightarrow \mathbb{R}_+^N$ is a vector valued function.

Remark 2.1. The equivalence of these two equations, and the triangle inequality for d_2 follows by Lemma B and Lemma A of Marton (2003).

Theorem 2.2. Let X, \hat{X}, \mathcal{M} and Γ be as in Theorem 2.1. Let $(\gamma_{i,j})_{i,j \leq n} := (\sqrt{\Gamma_{i,j}})_{i,j \leq n}$. Then for any distribution Q on Λ ,

$$d_2(P, Q) \leq \|\gamma\| \sqrt{s(\hat{X}) \cdot 2D(Q||P)}, \quad (2.14)$$

and

$$d_2(Q, P) \leq \|\gamma\| \sqrt{s(\hat{X}) \cdot 2D(Q||P)}. \quad (2.15)$$

Remark 2.2. This is a slight abuse of notation, because we also denote the spectral gap by γ , but since they will never appear in the same formula, they are easy to distinguish.

Corollary 2.5 (Talagrand's convex distance inequality). Let X, \hat{X}, \mathcal{M} and γ be as in Theorem 2.2. Let $A \subset \Lambda$, and $d_T(x, A)$ be the Talagrand distance of $x \in \Lambda$ from A :

$$d_T(x, A) := \sup_{\alpha: \Lambda \rightarrow \mathbb{R}_+^N, \sum \alpha_i^2 \leq 1} \inf_{y \in A} d_\alpha(x, y). \quad (2.16)$$

Then

$$\mathbb{E} \left(\exp \left(\frac{1}{4s(\hat{X})\|\gamma\|^2} d_T^2(X, A) \right) \right) \leq \frac{1}{P(A)}. \quad (2.17)$$

For the following result, we will need to define α -self-bounding functions (these are similar to self-bounding functions, see [Boucheron, Lugosi and Massart \(2009\)](#)).

Definition 7. Let $\Omega = \Omega_1 \times \dots \times \Omega_N$. Let $a, b \geq 0$.

1. We say that $f: \Omega \rightarrow \mathbb{R}$ is α -(a, b)-self-bounding if there is $\alpha: \Omega \rightarrow \mathbb{R}_+^N$ such that

$$(a) \ f(x) - f(y) \leq \sum_{i \leq N} \alpha_i(x) \mathbb{1}[x_i \neq y_i] \text{ for every } x, y \in \Omega.$$

$$(b) \ \alpha_i(x) \leq 1 \text{ for every } i \leq N, x \in \Omega.$$

$$(c) \ \sum_{i \leq N} \alpha_i(x) \leq af(x) + b.$$

2. We say that $f: \Omega \rightarrow \mathbb{R}$ is weakly α -(a, b)-self-bounding if there is $\alpha: \Omega \rightarrow \mathbb{R}_+^N$ such that

$$(a) \ f(x) - f(y) \leq \sum_{i \leq N} \alpha_i(x) \mathbb{1}[x_i \neq y_i] \text{ for every } x, y \in \Omega.$$

$$(b) \ \sum_{i \leq N} \alpha_i(x)^2 \leq af(x) + b.$$

Remark 2.3. It is easy to see that α -(a, b)-self-bounding functions are also weakly α -(a, b)-self-bounding. It is also easy to see that these are special cases of (a, b)-self-bounding and weakly (a, b)-self-bounding functions.

Theorem 2.3. Let X, \hat{X}, \mathcal{M} and γ be as in Theorem 2.2.

If $f: \Lambda \rightarrow \mathbb{R}$ is weakly α -(a, b)-self-bounding, then for every $\lambda > 0$,

$$\mathbb{E} \left(\exp \left(\lambda (f(X) - \mathbb{E}f(X)) - \frac{\lambda^2 \|\gamma\|^2 s(\hat{X}) a}{2} f(X) + \frac{\lambda^2 \|\gamma\|^2 s(\hat{X}) b}{2} \right) \right) \leq 1, \quad (2.18)$$

$$\mathbb{E} (\exp (-\lambda (f(X) - \mathbb{E}f(X)))) \leq \exp \left[\frac{\lambda^2 \|\gamma\|^2 (a \mathbb{E}f(X) + b)}{2} \right], \quad (2.19)$$

thus for every $t \geq 0$,

$$\mathbb{P}(f(X) \geq \mathbb{E}f(X) + t) \leq \exp \left(\frac{-t^2}{2\|\gamma\|^2 s(\hat{X}) (a\mathbb{E}f(X) + b + at)} \right), \quad (2.20)$$

$$\mathbb{P}(f(X) \leq \mathbb{E}f(X) + t) \leq \exp \left(\frac{-t^2}{2\|\gamma\|^2 s(\hat{X}) (a\mathbb{E}f(X) + b)} \right). \quad (2.21)$$

The following corollary is an improvement of Theorem 11.2 of [Dubhashi and Panconesi \(2009\)](#) (the constant is 2 times better in the independent case):

Corollary 2.6 (Method of non-uniformly bounded differences). *Let X, \hat{X}, \mathcal{M} and γ be as in Theorem 2.2. Suppose that there are $\alpha(x) := (\alpha_1(x), \dots, \alpha_N(x))$ real valued functions such that $f : \Lambda \rightarrow \mathbb{R}$ satisfies, for every x, y ,*

$$f(x) \leq f(y) + \sum_{i \leq N} \alpha_i(x) \mathbb{1}[x_i \neq y_i], \quad (2.22)$$

or

$$f(x) \geq f(y) - \sum_{i \leq N} \alpha_i(x) \mathbb{1}[x_i \neq y_i]. \quad (2.23)$$

Furthermore, suppose that there is a constant C such that for every $x \in \Lambda$,

$$\sum_{i=1}^N \alpha_i(x)^2 \leq C.$$

Then for every $t \geq 0$,

$$\mathbb{P}(f(X) - \mathbb{E}f(X) \geq t), \mathbb{P}(f(X) - \mathbb{E}f(X) \leq -t) \leq \exp \left(\frac{-t^2}{2\|\gamma\|^2 s(\hat{X}) C} \right). \quad (2.24)$$

The following is similar to Corollary 4 of [Samson \(2000\)](#):

Corollary 2.7 (Concentration for convex functions on a cube). *Let X, \hat{X}, \mathcal{M} and γ be as in Theorem 2.2. Additionally, suppose that $X_i, 1 \leq i \leq N$, take values in $[0, 1]$. Suppose that $f : [0, 1]^N \rightarrow \mathbb{R}$ is a 1-Euclidean Lipschitz, convex function. Then*

$$\mathbb{P}(f(X) - \mathbb{E}f(X) \geq t), \mathbb{P}(f(X) - \mathbb{E}f(X) \leq -t) \leq \exp \left(\frac{-t^2}{2\|\gamma\|^2 s(\hat{X})} \right). \quad (2.25)$$

Theorem 2.3 also implies concentration for supremum of positive valued empirical processes (similarly to Theorem 2 of [Samson \(2000\)](#)):

Corollary 2.8 (Concentration for positive valued empirical processes). *Let X, \hat{X}, \mathcal{M} and γ be as in Theorem 2.2. Let $(f_{i,j} : \Lambda_j \rightarrow [0, C])_{i \leq M, j \leq N}$ be a family of positive valued functions, bounded by C .*

Define

$$Z(x) := \sup_{j \leq M} \sum_{i \leq N} f_{i,j}(x_i) \text{ and } Z := Z(X). \quad (2.26)$$

Then $Z(x)/C$ is α -(1,0) self-bounding, and thus

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp \left(\frac{-t^2}{2\|\gamma\|^2 s(\hat{X}) C(\mathbb{E}Z + t)} \right), \quad (2.27)$$

$$\mathbb{P}(Z \leq \mathbb{E}Z - t) \leq \exp \left(\frac{-t^2}{2\|\gamma\|^2 s(\hat{X}) C\mathbb{E}Z} \right). \quad (2.28)$$

Remark 2.4. This formulation is analogous to the one in [Massart \(2000\)](#). The original formulation in the literature is

$$Z(x) := \sup_{f \in \mathcal{F}} \sum_{i \leq N} f(x_i)$$

for some countable set \mathcal{F} , our version is more general.

Theorem 2.4 (Bernstein inequality). *Let X , \hat{X} , \mathcal{M} and γ be as in Theorem 2.2. Suppose that $f_i : \Lambda_i \rightarrow [-C, C]$, and let*

$$V := \mathbb{E} \left(\sum_{i \leq N} f_i(X_i)^2 \right). \quad (2.29)$$

Let $S := \sum_{i \leq N} f_i(X_i)$, then for every $0 \leq \lambda \leq \frac{1}{2\sqrt{2}s(\hat{X})\|\gamma\|^2 C}$,

$$\log \mathbb{E}_P [\exp [\lambda(S - \mathbb{E}_P S)]] \leq \frac{2\|\gamma\|^2 s(\hat{X}) \mathbb{E}_P V \lambda^2}{1 - 2\sqrt{2}\|\gamma\|^2 s(\hat{X}) C \lambda}, \quad (2.30)$$

thus for every $t \geq 0$,

$$\begin{aligned} & \mathbb{P}(S \geq \mathbb{E}S + t), \mathbb{P}(S \leq \mathbb{E}S - t) \\ & \leq \exp \left(\frac{-t^2}{s(\hat{X})\|\gamma\|^2 (8V + 4\sqrt{2}C \cdot t)} \right). \end{aligned} \quad (2.31)$$

Remark 2.5. We do not require that $\mathbb{E}f_i(X_i) = 0$ for $i \leq N$.

Corollary 2.9 (Bernstein inequality for Markov chains). *First let $X = (X_1, \dots, X_N)$ be a (not necessarily time homogeneous) Markov chain taking values in some countable space $\Lambda := \Lambda_1 \times \dots \times \Lambda_N$. Suppose that $f_i : \Lambda_i \rightarrow [-C, C]$, $i \leq N$. Define $S := \sum_{i=1}^N f_i(X_i)$, and*

$$V := \mathbb{E} \left(\sum_{i=1}^N f_i(X_i)^2 \right). \quad (2.32)$$

Then for every $t \geq 0$,

$$\mathbb{P}(S \geq \mathbb{E}S + t), \mathbb{P}(S \leq \mathbb{E}S - t) \leq \exp \left(\frac{-t^2/\tau'_{\min}}{8V + 4\sqrt{2}Ct} \right). \quad (2.33)$$

Now suppose, in addition, that X is time homogeneous, irreducible, aperiodic, with stationary distribution π , and $\Lambda_1 = \dots = \Lambda_n = \Omega$. Suppose that $f : \Omega \rightarrow [-C, C]$. Let $t_0 \geq 0$ (“burn-in time”), and $Z := \frac{\sum_{i=t_0+1}^N f(X_i)}{N-t_0}$. In this case,

$$V = \mathbb{E} \left(\sum_{i=1}^N f(X_i)^2 \right). \quad (2.34)$$

Then for every $t \geq 0$,

$$\begin{aligned} & \mathbb{P} \left(Z \geq \mathbb{E}_\pi(f) + \frac{2C\eta_{\min}(t_0)}{N-t_0} + t \right), \quad \mathbb{P} \left(Z \leq \mathbb{E}_\pi(f) - \frac{2C\eta_{\min}(t_0)}{N-t_0} - t \right) \\ & \leq \exp \left(\frac{-t^2(N-t_0)^2/\tau'_{\min}}{8V + 4\sqrt{2} \cdot (N-t_0)Ct} \right). \end{aligned} \quad (2.35)$$

The following result is similar to Theorem 3 of [Samson \(2000\)](#):

Theorem 2.5 (Weak version of Talagrand’s suprema of empirical processes inequality). *Let X, \hat{X}, \mathcal{M} and γ be as in Theorem 2.2.*

Let $(f_{i,j} : \Lambda_j \rightarrow [-C, C])_{i \leq M, j \leq N}$ be a family of functions, bounded by C .

Define

$$Z(x) := \sup_{j \leq M} \sum_{i \leq N} f_{i,j}(x_i), \text{ and } Z := Z(X), \quad (2.36)$$

define

$$G(\lambda) := \log \mathbb{E} e^{\lambda(Z - \mathbb{E}Z)}, \quad (2.37)$$

and let

$$W := \mathbb{E} \left(\sum_{i \leq N} \max_{j \leq M} f_{i,j}(X_i)^2 \right). \quad (2.38)$$

Then for every $\lambda > 0$,

$$G(\lambda), G(-\lambda) \leq \frac{4s(\hat{X})\|\gamma\|^2 W \lambda^2}{1 - 2\sqrt{2}\|\gamma\|^2 |\lambda| C}, \quad (2.39)$$

thus for every $t \geq 0$,

$$\begin{aligned} & \mathbb{P}(Z \geq \mathbb{E}Z + t), \mathbb{P}(Z \leq \mathbb{E}Z - t) \\ & \leq \exp \left(\frac{-t^2}{s(\hat{X})\|\gamma\|^2 (16W + 4\sqrt{2}C \cdot t)} \right). \end{aligned} \quad (2.40)$$

The same inequalities hold with the definition

$$Z(x) := \left| \sup_{j \leq M} \sum_{i \leq N} f_{i,j}(x_i) \right|. \quad (2.41)$$

Our next result is an extension of Theorem 3 and 3’ of [Marton \(2003\)](#):

Theorem 2.6. Let X, \hat{X}, \mathcal{M} and γ be as in Theorem 2.2.

Suppose that $f : \Lambda \rightarrow \mathbb{R}$ satisfies one of the following:

Condition 1. There are functions $\alpha_i : \Lambda \rightarrow \mathbb{R}_+, i \leq N$, such that for any $x, y \in \Lambda$,

$$f(x) - f(y) \leq \sum_{i=1}^N \alpha_i(x) \mathbb{1}[x_i \neq y_i].$$

Condition 2. There are functions $\alpha_i : \Lambda \rightarrow \mathbb{R}_+, \beta_i : \Lambda \rightarrow \mathbb{R}_+, i \leq N$, such that for any $x, y \in \Lambda$,

$$f(x) - f(y) \leq \sum_{i=1}^N (\alpha_i(x) + \beta_i(y)) \mathbb{1}[x_i \neq y_i].$$

Let us denote

$$\begin{aligned} F(\lambda) &:= \mathbb{E} e^{\lambda(f(X) - \mathbb{E}f(X))}, \\ G(\lambda) &:= \log F(\lambda), \\ V_\alpha &:= \mathbb{E} \sum_{i=1}^N \alpha_i^2(X), \\ V_\beta &:= \mathbb{E} \sum_{i=1}^N \beta_i^2(X), \\ g_\alpha(\tau) &:= \log \mathbb{E} e^{\tau \sum_{i=1}^N \alpha_i^2(X)}, \\ g_\beta(\tau) &:= \log \mathbb{E} e^{\tau \sum_{i=1}^N \beta_i^2(X)}. \end{aligned}$$

If f satisfies Condition 1, then for $\lambda > 0$,

$$G(\lambda) \leq \min_{\tau > 2\lambda^2 \|\gamma\|^2} \frac{2\lambda^2 \|\gamma\|^2}{\tau - 2\lambda^2 \|\gamma\|^2} g_\alpha(s(\hat{X}) \cdot \tau), \quad (2.42)$$

and

$$G(-\lambda) \leq 2s(\hat{X})\lambda^2 \|\gamma\|^2 V_\alpha. \quad (2.43)$$

If f satisfies Condition 2, then for $\lambda > 0$,

$$G(\lambda) \leq \min_{\tau > 4\lambda^2 \|\gamma\|^2} \frac{4\lambda^2 \|\gamma\|^2}{\tau - 4\lambda^2 \|\gamma\|^2} \left(g_\alpha(s(\hat{X})\tau) + s(\hat{X})\tau V_\beta \right), \quad (2.44)$$

and

$$G(-\lambda) \leq \min_{\tau > 4\lambda^2 \|\gamma\|^2} \frac{4\lambda^2 \|\gamma\|^2}{\tau - 4\lambda^2 \|\gamma\|^2} \left(g_\beta(s(\hat{X})\tau) + s(\hat{X})\tau V_\alpha \right). \quad (2.45)$$

Remark 2.6. This result is quite powerful, since all of the previous inequalities follow from it (with slightly worse constants).

2.2. Results by spectral methods

In this section, we give some concentration inequalities for empirical averages, using spectral methods. For finite (or countable) state reversible chains, the sharp version of Hoeffding's inequality has the following form (here we have adapted it to work for non-stationary initial distribution too):

Theorem 2.7 (Theorem 1 of [León and Perron \(2004\)](#)). *Let $X = (X_1, \dots, X_N)$ be a time homogeneous, reversible, irreducible, aperiodic Markov chain taking values in some finite state space Ω , with stationary distribution π . Let λ be the second largest eigenvalue of P ($\lambda = 1 - \gamma$), and $f : \Omega \rightarrow [a, b]$. Denote $S := \sum_{i=1}^N f(X_i)$, and let $\lambda_0 = \max(0, \lambda)$. Suppose that $X_1 \sim \pi$, then*

$$\mathbb{P} \left[\frac{S}{N} \geq \mathbb{E}_\pi f + t \right], \mathbb{P} \left[\frac{S}{N} \leq \mathbb{E}_\pi f - t \right] \leq \exp \left(-2 \frac{1 - \lambda_0}{1 + \lambda_0} N t^2 / (b - a)^2 \right). \quad (2.46)$$

For arbitrary initial distribution, denote $Z := \frac{1}{N - t_0} \sum_{i=t_0+1}^N f(X_i)$, then we have

$$\mathbb{P} [Z \geq \mathbb{E}_\pi f + t], \mathbb{P} [Z \leq \mathbb{E}_\pi f - t] \leq \exp \left(-2 \frac{1 - \lambda_0}{1 + \lambda_0} (N - t_0) t^2 / (b - a)^2 \right) + \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor}. \quad (2.47)$$

Remark 2.7. The proof of (2.47) follows by the same argument that we use in the proof of Corollary 2.10.

Now we present a Bernstein-type result for finite state reversible chains, which is based on the proof of Theorem 1.1. of [Lezaud \(1998a\)](#):

Corollary 2.10 (Bernstein inequality for reversible Markov chains). *Let $X = (X_1, \dots, X_N)$ be a time homogeneous, reversible, irreducible, aperiodic Markov chain taking values in some finite state space Ω , with stationary distribution π , spectral gap γ , and mixing time $t_{\text{mix}}(\epsilon)$ for some $0 \leq \epsilon < 1$. Suppose that $f : \Omega \rightarrow [-C, C]$ with $\mathbb{E}_\pi f = 0$, and denote $V_f := \text{Var}_\pi(f)$.*

Let $t_0 \geq 0$ (“burn-in time”), define $Z := \frac{\sum_{i=t_0+1}^N f(X_i)}{N - t_0}$, and let

$$h(x) := \frac{1}{2} \left(\sqrt{1 + x} - (1 - x/2) \right), \quad (2.48)$$

then for $t \geq 0$,

$$\begin{aligned} & \mathbb{P} [Z - \mathbb{E}_\pi f \geq t], \mathbb{P} [Z - \mathbb{E}_\pi f \leq -t] \\ & \leq e^{\gamma/5} \exp \left[-\frac{(N - t_0)t^2\gamma}{4V_f + 4h(5Ct/V_f)} \right] + \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor} \end{aligned} \quad (2.49)$$

$$\leq e^{\gamma/5} \exp \left[-\frac{(N - t_0)t^2\gamma}{4V_f + 10C \cdot t} \right] + \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor}. \quad (2.50)$$

Define the asymptotic variance, σ^2 , as

$$\sigma^2 := \lim_{N \rightarrow \infty} \frac{1}{N} \text{Var}_\pi (f(X_1) + \dots + f(X_N)), \quad (2.51)$$

then the following bounds hold:

$$\mathbb{P}[Z - \mathbb{E}_\pi f \geq t], \mathbb{P}[Z - \mathbb{E}_\pi f \leq -t] \leq \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor} + e^{\gamma/5} \exp \left[-(N - t_0) \cdot \left(\frac{\sqrt{\left(\sigma^2 + \frac{5}{\gamma}tC\right)^2 + 4\sigma^2 K' t C} - \left(\sigma^2 + \frac{5}{\gamma}tC\right)}{2\sigma^2 K'} \right) \cdot \frac{t}{C} \right], \quad (2.52)$$

with $K' := \frac{10V_f}{\gamma^2 \sigma^2} - \frac{5}{\gamma}$.

Remark 2.8. We got rid of N_q in Theorem 1.1, and thus this form of the bound is more useful for practical applications.

Theorem 3.3. of [Lezaud \(1998a\)](#) (see also Theorem 2.1 in [Lezaud \(1998b\)](#)) generalizes this bound to non-reversible chains, with constants in the exponent depending on the spectral gap of the multiplicative symmetrization $K := P^*P$, where P^* is the adjoint of P in $\ell^2(\pi)$. The weakness of this approach is that the spectral gap of K can be very small, or even zero, and it is not necessarily related to the mixing time of the chain. We propose the following improved version, which settles this difficulty:

Theorem 2.8 (Bernstein inequality for non-reversible Markov chains). *Define the pseudo spectral gap for an irreducible, aperiodic P with stationary distribution π as*

$$\gamma_{\text{ps}} := \sup_{k \geq 1} \frac{\gamma((P^*)^k P^k)}{k}. \quad (2.53)$$

With the notations of [Corollary 2.10](#), we have, for $t \geq 0$,

$$\begin{aligned} & \mathbb{P}[Z - \mathbb{E}_\pi f \geq t], \mathbb{P}[Z - \mathbb{E}_\pi f \leq -t] \\ & \leq \exp \left[-\frac{(N - t_0)t^2 \gamma_{\text{ps}}}{8V_f + 8h(5Ct/V_f)} \right] + \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor} \end{aligned} \quad (2.54)$$

$$\leq \exp \left[-\frac{(N - t_0)t^2 \gamma_{\text{ps}}}{8V_f + 20C \cdot t} \right] + \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor}. \quad (2.55)$$

Remark 2.9. For $k \gg t_{\text{mix}}$, $P^k \approx \lim_{t \rightarrow \infty} P^t$, and $\gamma((\lim_{t \rightarrow \infty} P^t)^* \lim_{t \rightarrow \infty} P^t) = 1$, so γ_{ps} can not be much smaller than $1/t_{\text{mix}}$.

2.3. Results for continuous time chains

Our next two results are based on [Corollaries 2.4 and 2.9](#), and show concentration inequalities for (not necessarily reversible) continuous time Markov chains (with countable state space). The proof of these are left to the reader as exercise (they can be done using the same technique as in the proof of Theorem 3.4. on page 858 of [Lezaud \(1998a\)](#)).

Corollary 2.11 (Hoeffding inequality for continuous time Markov chains). *Let $(X_s)_{s \geq 0}$ be a time homogeneous, continuous time Markov chain taking values in some countable space*

Ω , stationary distribution π , and mixing time $t_{\text{mix}}^{\text{cont}}$. Let $f : \Omega \rightarrow [a, b]$. Let $t_0 \geq 0$ (“burn-in time”), denote

$$Z := \frac{1}{T - t_0} \int_{s=t_0}^T f(X_s) ds. \quad (2.56)$$

Then for every $t \geq 0$,

$$\begin{aligned} & \mathbb{P} \left(Z \geq \mathbb{E}_\pi(f) + \frac{(b-a)\eta_{\min}^{\text{cont}}(t_0)}{T-t_0} + t \right), \mathbb{P} \left(Z \leq \mathbb{E}_\pi(f) - \frac{(b-a)\eta_{\min}^{\text{cont}}(t_0)}{T-t_0} - t \right) \\ & \leq \exp \left(\frac{-2(T-t_0)t^2}{(b-a)^2\tau_{\min}^{\text{cont}}} \right). \end{aligned} \quad (2.57)$$

Corollary 2.12 (Bernstein inequality for continuous time Markov chains). *Let $(X_s)_{s \geq 0}$ be as in Corollary 2.11. Let $f : \Omega \rightarrow [-C, C]$, and Z as in (2.56).*

Denote

$$V := \mathbb{E} \left(\int_{s=t_0}^T f(X_s)^2 ds \right). \quad (2.58)$$

Then for every $t \geq 0$,

$$\begin{aligned} & \mathbb{P} \left(Z \geq \mathbb{E}_\pi(f) + \frac{2C\eta_{\min}^{\text{cont}}(t_0)}{T-t_0} + t \right), \mathbb{P} \left(Z \leq \mathbb{E}_\pi(f) - \frac{2C\eta_{\min}^{\text{cont}}(t_0)}{T-t_0} - t \right) \\ & \leq \exp \left(\frac{-t^2(T-t_0)^2/\tau_{\min}^{\text{cont}}}{8V + 4\sqrt{2}(T-t_0)Ct} \right). \end{aligned} \quad (2.59)$$

As a comparison, the following theorem is the main result of [Lezaud \(2001\)](#):

Theorem 2.9 (Theorem 1.1. of [Lezaud \(2001\)](#)). *Let P_t be an ergodic Markov semigroup with invariant probability measure π . Assume that its infinitesimal generator L has as simple isolated eigenvalue $\lambda = 0$ and that the initial distribution q has a $L^2(\pi)$ density relatively to the measure π . Then, for all $f \in D_2(L)$ such that $\pi(f) = 0$, $\|f\|_\infty \leq a$, for all $t > 0$ and $T > 0$,*

$$\mathbb{P}_q(T^{-1}S_T \geq t) \leq N_q \exp \left\{ - \frac{2Tt^2}{\sigma^2 \left(1 + \sqrt{1 + 4at/(\gamma\sigma^2)} \right)^2} \right\}, \quad (2.60)$$

with $S_T := \int_0^T f(X_s) ds$, $\sigma^2 := \lim_{T \rightarrow \infty} T^{-1} \text{Var}_\pi(S_T)$, γ is the spectral gap of $(L + L^*)/2$, and N_q is the $L^2(\pi)$ norm of the density of q related to the stationary distribution π .

3. Applications

3.1. Coin tossing

The reader might think that independent Bernoulli trials is a good model for coin tossing.

We disagree. In a famous paper, [Diaconis, Holmes and Montgomery \(2007\)](#), it was shown that it is slightly more likely for the coin to come up on the same side as it was at the beginning.

We have made our experiments with a Singapore 50 cent coin, and tossed it up 40-50 cm high. Our results for 1200 coin tosses (1 corresponds to heads, and 0 to tails):

We have used this sequence to estimate the variance of

$$S_n := X_1 + \dots + X_n.$$

$$m_s = 0.5333 \text{ and } V_s = 16.5747. \quad (3.1)$$
$$T := \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}.$$

We estimate these probabilities by the counts of 00, 01, 10, 11, which we denote by $\#00$, ect. Thus we get the estimate

$$\hat{T} := \begin{pmatrix} \frac{\#00}{\#00+\#01} & \frac{\#01}{\#00+\#01} \\ \frac{\#10}{\#10+\#11} & \frac{\#11}{\#10+\#11} \end{pmatrix} \approx \begin{pmatrix} 0.6071 & 0.3929 \\ 0.3443 & 0.6557 \end{pmatrix}.$$

Let X_1, \dots, X_{1200} be a Markov chain with this transition matrix, started from 1. Since we do not have a closed formula for V_s for this model, we just ran a long computer simulation (100000 runs, 1200 steps in each), which gave us that the expected value of this variance is approximately 16.8, which is very close to what we have observed. This means that the Markov chain model describes the real situation better than i.i.d. Bernoulli trials.

For such a Markov chain, Corollary 2.4, or Theorem 2.8 can be applied to bound the deviation probabilities of S_n from its expected value.

The reader should not think of this as an isolated example, Markov chain models have been successfully applied to many real life situations. For an example about basketball gambling which outperformed the models of the bookmakers, see [Kvam and Sokol \(2006\)](#).

3.2. Error analysis for MCMC

MCMC methods have a huge literature. They are used, amongst other things, for simulating distributions arising from statistical physics, approximate counting in combinatorial structures, approximate integration, stochastic optimization (simulated annealing), ect. Our favourite review papers are [Diaconis \(2009\)](#) and [Jerrum and Sinclair \(1996\)](#).

Let X_1, \dots, X_N be a time homogeneous Markov chain, taking values in $\Lambda = \Omega^N$, with stationary distribution π . We may be interested in computing the expectation $\mathbb{E}_\pi f$ for some function $f : \Omega \rightarrow \mathbb{R}$. This can be approximated by the average

$$\mathbb{E}_\pi f \approx \frac{f(X_1) + \dots + f(X_N)}{N}.$$

A natural question to ask is how large N should be so that this approximation is good, i.e. how long should we run our simulation.

This problem have been extensively studied in [Lezaud \(1998b\)](#) for reversible Markov chains with finite/general state space, and reversible Markov processes with finite/general state space, with Bernstein-type results proven in all situations, that are roughly $1/\gamma$ times weaker than in the independent case. However, these results involve a constant N_q that depends on the initial state of the chain, which may be difficult to compute in practice, and the results for non-reversible chains are not satisfactory.

Given the mixing time $t_{\text{mix}}(\epsilon)$, and the spectral gap γ of the chain, Corollary 2.4, Corollary 2.9, and Corollary 2.10 gives bounds on the deviation $\frac{f(X_1) + \dots + f(X_N)}{N} - \mathbb{E}_\pi f$.

Finding bounds on the spectral gap and the mixing time has a large literature, we refer the reader to [Levin, Peres and Wilmer \(2009\)](#), and [Lovász and Winkler \(1998\)](#).

One example of such a bound is Theorem 3 in [Bubley et al. \(1997\)](#): under the Dobrushin uniqueness condition, i.e. if the maximum column sum of the Dobrushin matrix is $\alpha < 1$, the Gibbs sampler Markov chain of a statistical physical model has mixing time

$$t_{\text{mix}}(\epsilon) \leq \lceil n \log(n\epsilon^{-1}) / (1 - \alpha) \rceil. \quad (3.2)$$

This bound works for many statistical physical models (Curie-Weiss, Ising, Potts, ect.) at sufficiently high temperature.

For a bound on the coefficients of the Dobrushin matrix, see [Chatterjee \(2005\)](#), page 79, Lemma 4.4.

For more examples and simulation results, we refer the reader to [Gyori and Paulin \(2012\)](#).

3.3. m -dependence

We say that X_1, \dots, X_N are m -dependent random variables if for each $1 \leq i \leq N - m$, (X_1, \dots, X_i) and (X_{i+m}, \dots, X_N) are independent.

For this dependence structure, we can define $n := \lceil \frac{N}{m} \rceil$,

$$\hat{X}_1 := (X_1, \dots, X_m), \dots, \hat{X}_N := (X_{(n-1)m+1}, \dots, X_N).$$

For \hat{X} , we construct a Marton coupling \mathcal{M} :

$$\mathcal{M}^i \left[\hat{X}_{\geq i+1}^n \sim \hat{P}_{\geq i+1}^n(\cdot | \hat{x}_{\leq i}), \hat{X}'_{\geq i+1}^n \sim P_{\geq i+1}^n(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i) \right]$$

is constructed by first defining $\hat{X}_{\geq i+2}^n = \hat{X}'_{\geq i+2}^n$, with distribution $\hat{P}_{\geq i+2}^n$ (here we use the m -dependence condition), and then defining \hat{X}_{i+1} and \hat{X}'_{i+1} conditionally on these two. Therefore, it is clear that the mixing matrix for \mathcal{M} satisfies

$$\Gamma = (\Gamma_{i,j})_{i,j \leq n} \leq \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \quad (3.3)$$

so we can see that $\|\Gamma\| \leq 2$ and $\gamma \leq 2$, so our theorems hold under this condition, with $s(\hat{X}) = m$. Thus the constants in the exponents are about $4m$ times worse than in the independent case.

We finish this section with the following “metatheorem”:

Metatheorem 3.1. *Suppose that X_1, \dots, X_N are dependent random variables that can be put in a sequence with a typical range of dependence m . Then the concentration inequalities hold with constants cm times weaker than in the independent case, for some constant c (independent of N, m).*

Proof. Define \hat{X} as in the m - dependent case, and then construct the Marton coupling \mathcal{M} for \hat{X} . \square

3.4. Independent permutations and Bernoulli variables with fixed sum

Let X_1, \dots, X_n be

1. Uniformly chosen random permutations of $1, \dots, n$, or
2. 0,1 valued random variables with $\sum_{i=1}^n X_i = k$ for some $0 \leq k \leq n$, and uniformly distributed among the $\binom{n}{k}$ possibilities.

Then

$$\mathcal{M}^i \left[\hat{X}_{\geq i+1}^n \sim \hat{P}_{\geq i+1}^n(\cdot | \hat{x}_{\leq i}), \hat{X}'_{\geq i+1}^n \sim P_{\geq i+1}^n(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i) \right]$$

is constructed by letting I be distributed uniformly on $i+1, \dots, n$, choosing $X_I = x'_i$, $X'_I = x_i$, and then setting the rest of $X_{\geq i+1}^n$ and $X'_{\geq i+1}^n$ the same.

For such a coupling, we have the coupling matrix

$$\Gamma = (\Gamma_{i,j})_{i,j \leq n} \leq \begin{pmatrix} 1 & \frac{1}{n-1} & \frac{1}{n-1} & \frac{1}{n-1} & \frac{1}{n-1} & \frac{1}{n-1} & \cdots \\ 0 & 1 & \frac{1}{n-2} & \frac{1}{n-2} & \frac{1}{n-2} & \frac{1}{n-2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}. \quad (3.4)$$

This means that for 1-Hamming Lipschitz functions, i.e. c weighted Hamming Lipschitz functions with $c = (1, \dots, 1)$, we have $\|\Gamma \cdot c\|^2 = 4$, so by Theorem 2.1, the McDiarmid and Hoeffding inequalities hold with constant 4 times worse than in the independent case.

For permutations, a much stronger result, Talagrand's convex distance inequality, with constant 4 times weaker than in the independent case, was proven in Section 5 of Talagrand (1995). This was further developed in McDiarmid (2002). For an overview, see Section 8.2 of Ledoux (2001). See also Chatterjee (2007) for a concentration inequality in the setting of the combinatorial central limit theorem.

Unfortunately, $\|\Gamma\| \sim \sqrt{\log(n)}$ and $\|\gamma\| \sim \sqrt{n}$, so we can not recover these results.

3.5. Hidden Markov chains

Concentration inequalities for Hidden Markov chains have been investigated in Kontorovich (2006), see also Kontorovich (2007), Section 4.1.4.

Let $\tilde{X}_1, \dots, \tilde{X}_N$ be a Markov chain (not necessarily homogeneous) taking values in a discrete set $\tilde{\Lambda} = \tilde{\Lambda}_1 \times \dots \times \tilde{\Lambda}_N$, with distribution \tilde{P} .

Let X_1, \dots, X_N be random variables taking values in the discrete space $\Lambda = \Lambda_1 \times \dots \times \Lambda_N$ such that the joint distribution of (\tilde{X}, X) is given by

$$H(\tilde{x}, x) := \tilde{P}(\tilde{x}) \cdot \prod_{i=1}^n P_i(x_i | \tilde{x}_i),$$

i.e. X_i are conditionally independent given \tilde{X} . Then we call X_1, \dots, X_N a hidden Markov chain.

The following result (an extension of Theorem 4.1.4 of Kontorovich (2007) to our setting) shows that the concentration properties of a hidden Markov chain are completely determined by the concentration properties of the underlying chain.

Proposition 3.1. *Let*

$$\begin{aligned} \hat{\tilde{X}} &:= (\hat{\tilde{X}}_1, \dots, \hat{\tilde{X}}_n) := \\ &\left((\tilde{X}_1, \dots, \tilde{X}_{i_1}), (\tilde{X}_{i_1+1}, \dots, \tilde{X}_{i_2}), \dots, (\tilde{X}_{i_{n-1}+1}, \dots, \tilde{X}_N) \right) \\ \hat{X} &:= (\hat{X}_1, \dots, \hat{X}_n) := \\ &((X_1, \dots, X_{i_1}), (X_{i_1+1}, \dots, X_{i_2}), \dots, (X_{i_{n-1}+1}, \dots, X_N)) \end{aligned}$$

be partitions of \tilde{X} and X . Suppose that $\tilde{\mathcal{M}}$ is a Marton coupling for \tilde{X} , with mixing matrix $\tilde{\Gamma}$, then there is a Marton coupling \mathcal{M} for \hat{X} with mixing matrix $\Gamma \leq \tilde{\Gamma}$ (in each element).

Proof. Suppose first that $\hat{X} = X$, then $n = N$ and $s(\hat{X}) = 1$ (the general case is similar).

We are given

$$\tilde{\mathcal{M}}^i \left[\tilde{X}_{\geq i+1}^n \sim \tilde{P}_{\geq i+1}^n(\cdot | \tilde{x}_{\leq i}), \tilde{X}'_{\geq i+1}^n \sim \tilde{P}_{\geq i+1}^n(\cdot | \tilde{x}_{\leq i-1}, \tilde{x}'_i) \right],$$

and need to construct

$$\mathcal{M}^i \left[X_{\geq i+1}^n \sim P_{\geq i+1}^n(\cdot | x_{\leq i}), X'_{\geq i+1}^n \sim P_{\geq i+1}^n(\cdot | x_{\leq i-1}, x'_i) \right].$$

This can be done by first defining a coupling

$$\begin{aligned} \pi \left((\tilde{X}_{\geq i+1}^n, \tilde{X}'_{\geq i+1}^n) \sim \tilde{\mathcal{M}}^i(\cdot | \tilde{x}_{\leq i}, \tilde{x}'_i), X_{\geq i+1}^n \sim H(\cdot | \tilde{x}_{\leq i}, \tilde{X}_{\geq i+1}^n), X'_{\geq i+1}^n \right. \\ \left. \sim H(\cdot | \tilde{x}_{\leq i-1}, \tilde{x}'_i, \tilde{X}'_{\geq i+1}^n) \right), \end{aligned}$$

satisfying that given $\tilde{X}_{\geq i+1}^n$ and $\tilde{X}'_{\geq i+1}^n$, $(X_{i+1}, X'_{i+1}) \dots, (X_n, X'_n)$ are independent, with (X_j, X'_j) distributed as the maximal coupling of the distributions $P_j(\cdot | X_j)$ and $P_j(\cdot | X'_j)$.

One can see that, by the Markov property, marginal distribution of $X_{\geq i+1}^n$ and $X'_{\geq i+1}^n$ only depends on x_i and x'_i and does not depends on $x_{\leq i-1}$.

Therefore, we can construct \mathcal{M}^i by first defining $(\tilde{X}_i, \tilde{X}'_i)$ as the maximal coupling of $\tilde{P}_i(\cdot | x_{\leq i})$ and $\tilde{P}_i(\cdot | x_{\leq i-1}, x'_i)$, and then defining the rest of it as in π , given $\tilde{X}_i, \tilde{X}'_i$. Finally, it is easy to check that $\Gamma \leq \tilde{\Gamma}$.

Note that the Markov property is necessary for this proof to work, see [Kontorovich \(2007\)](#), page 36 for a counterexample when \tilde{X} is not Markov. □

3.6. Random walks on weighted graphs

We adapt the notations of [Gillman \(1998\)](#). Let $G = (V, E)$ be a connected undirected graph, with each edge $\{x, y\}$ in \mathbb{E} having weight w_{xy} . Let $w_x := \sum_{y: \{x, y\} \in \mathbb{E}} w_{xy}$ be the weight of x .

Then a random walk on G is equivalent to a time reversible Markov chain with transition matrix $T := (p_{xy})_{xy \in \mathbb{E}}$, with

$$p_{xy} = \begin{cases} \frac{w_{xy}}{w_x} & \text{if } \{x, y\} \in \mathbb{E}, \\ 0 & \text{if not.} \end{cases}$$

We denote the eigenvalues of T by $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{|V|}$. We denote by $\gamma := 1 - \lambda_2$ the *eigenvalue gap*, this is strictly positive for connected graphs. The stationary distribution of the walk is denoted by

$$\pi(x) = \frac{w_x}{\sum_{y \in V} w_y}.$$

The main result of [Gillman \(1998\)](#) is the following theorem (see also [Kahale \(1997\)](#) for a sharper version):

Theorem 3.1 (Theorem 2.1 of [Gillman \(1998\)](#)). *Consider the random walk on a weighted graph $G = (V, E)$ with initial distribution q . Let $A \subseteq V$. Let A_n be the number of visits to A in n steps. Let $\frac{q}{\sqrt{\pi}}$ denote the vector $\frac{q}{\sqrt{\pi}}(x) = \frac{q(x)}{\sqrt{\pi(x)}}$, and $N_q = \left\| \frac{q}{\sqrt{\pi}} \right\|_2$. For any $t \geq 0$,*

$$\mathbb{P}(A_n - n\pi(A) \geq t), \mathbb{P}(A_n - n\pi(A) \leq -t) \leq (1 + t\varepsilon/(10n)) N_q e^{-t^2\gamma/(20n)}, \quad (3.5)$$

and the same bound holds for the lower tail.

Remark 3.1. A similar bound can be deduced from Theorem 1.1. of [Lezaud \(1998a\)](#), and thus by Corollary 2.10. See also Theorem 2.7.

This theorem is very useful for probability amplification, here we briefly review [Dubhashi and Panconesi \(2009\)](#), Section 3.5.3.

Suppose that one has a random algorithm, computing a function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, which takes in $x \in \{0, 1\}^n$ and an n long sequence of random bits r , and gives the result $A(x, r)$. Suppose that the probability of correct evaluation is bounded away from 0, lets say

$$\mathbb{P}(A(x, r) = f(x)) \geq 3/4.$$

Then the goal of probability amplification is to increase this success probability. This is can be done by running the algorithm k times for k independent r , and then the chance of having at least half of the results the same is, by Hoeffding inequality for independent variables, smaller than $e^{-k/8}$.

Such a direct method would use nk random bits.

We can get this results using fewer bits, the following way. First, let $G = (\{0, 1\}^n, E)$ be connected d -regular undirected graph, more precisely, an expander graph (the reader can consult [Hoory, Linial and Wigderson \(2006\)](#), [Tao \(2010\)](#), and [Kowalski \(2011\)](#) for a review). These graphs were popularized by the papers [Ajtai, Komlós and Szemerédi \(1983\)](#) and [Ajtai, Komlós and Szemerédi \(1987\)](#).

For our purposes, it suffices that there exists graphs that the eigenvalue gap of the associated “uniformly” weighted random walk is roughly $\frac{1}{\sqrt{d}}$. Let G be such a graph.

We take r_1 to be uniformly distributed in $\{0, 1\}^n$ (which is the stationary distribution π), and r_2, \dots, r_l be a random walk on G . Then using Theorem 3.1, we can prove that the chance of having less than half of $A(x, r_1), \dots, A(x, r_l)$ be correctly evaluated, is less than $e^{-cl\varepsilon}$ for some universal constant c . Since generating r_1, \dots, r_l only takes $n + l \log_2 d$ bits, and $\varepsilon \geq \frac{1}{\sqrt{d}}$, we can see that choosing $l = \sqrt{dk}/c$ gives the same precision as k independent n long sequences, and takes considerably less, $n + k \log_2 d \sqrt{d}/c$ random bits.

A natural question: how is this setting related to our theorems?

Let us denote the mixing time of the random walk on G by t_{mix} , then we can write $A_n = \sum_{i=1}^n \mathbb{1}[X_i \in A]$, and use Corollary 2.4, to get that the concentration of A_n around its mean is about t_{mix} times worse than in the independent case.

It is easy to see that in k steps, we can visit less than d^k vertices, and will be far away from the uniform stationary distribution in total variation distance unless $k \geq \log_d |V| = n \log_d 2$. Therefore $t_{\text{mix}} \geq n \log_d 2$. This means that Corollary 2.4, in this case, is much weaker than Theorem 3.1 and Corollary 2.10.

However, we seriously doubt that such an inequality could hold for more general functions, for example sums of the form $\sum_{i=1}^N f_i(X_i)$, or Hamming-Lipschitz functions.

Therefore we consider this as an example of “superconcentration”: for empirical sums of the form $\sum_{i=1}^N f(X_i)$, much stronger concentration occurs than for arbitrary functions.

Beyond probability amplification, this technique can be also used to evaluate the expectation of any $f : \{0, 1\}^n \rightarrow [a, b]$, real valued function by approximating it by

$$\mathbb{E}f(X) \approx \frac{\sum_{i=1}^N f(X_i)}{N}, \quad (3.6)$$

X and X_1 being uniformly distributed in $\{0, 1\}^n$, and $\{X_i\}_{2 \leq i \leq N}$ being the random walk on the expander G . This approximation uses only $n + N \log_2 d$ random bits (instead of Nn bits by the independent sampling), and its precision can be estimated by Theorems 2.7, and 2.10 (the constants are roughly \sqrt{d} times worse than in the independent case).

4. Open problems

Since most of our concentration results are roughly t_{mix} times weaker than in the independent case, the following questions naturally arise:

1. (Talagrand’s suprema of empirical processes) It would be interesting to prove the complete version of Talagrand’s suprema for empirical processes inequality, i.e. improve Theorem 2.5 by replacing W with

$$V := \mathbb{E} \left(\max_{j \leq M} \sum_{i \leq N} f_{i,j}(X_i)^2 \right). \quad (4.1)$$

One could try to further bound V with $\max_{j \leq M} \mathbb{E} \left(\sum_{i \leq N} f_{i,j}(X_i)^2 \right)$ and $\mathbb{E} \left(\max_{j \leq M} \left| \sum_{i \leq N} f_{i,j}(X_i) \right| \right)$, as it is done in the independent case (the proof for this result on page 141 of [Ledoux \(2001\)](#), and page 112 of [Ledoux and Talagrand \(1991\)](#), breaks down in the case of dependence). For an elegant proof of these results in the independent case using the entropy method, see 169-170 of [Massart \(2007\)](#).

The reader could approach this problem by further developing Theorem 2.6, or by adapting Talagrand’s q point method to the dependent case, see [Talagrand \(1996\)](#), [Dembo \(1997\)](#).

2. (Unbounded random variables) Lemma 5.5. of [Vershynin \(2010\)](#) shows that three natural definitions of subgaussian random variables (tail bound, moment bound, subexponential moment) are in fact equivalent.

Definition 5.7. of [Vershynin \(2010\)](#) defines the ψ_2 norm of a real valued random variable X as

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p}. \quad (4.2)$$

For bounded variables, we have $\|X\|_{\psi_2} \leq \|X\|_{\infty}$.

Proposition 5.10 of [Vershynin \(2010\)](#) gives a Chernoff-Hoeffding type inequality:

Proposition 4.1. *Let X_1, \dots, X_N be independent, centered, subgaussian random variables, and let $K := \max_i \|X_i\|_{\psi_2}$. Then for every $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P} \left(\left| \sum_{i=1}^N a_i X_i \right| \geq t \right) \leq e \cdot \exp \left(-\frac{ct^2}{K^2 \cdot \sum_{i \leq N} a_i^2} \right), \quad (4.3)$$

where $c > 0$ is an absolute constant.

Theorem A.7.1 of [Talagrand \(2011\)](#) is the following unbounded version of Bernstein's inequality:

Theorem 4.1. *Let X_1, \dots, X_N be iid centered random variables, distributed like X with $\mathbb{E}X = 0$. Assume that*

$$\mathbb{E} \exp \frac{|X|}{A} \leq 2.$$

Then for all $t > 0$ we have

$$\mathbb{P} \left(\sum_{i \leq N} X_i \geq t \right) \leq \exp \left(-\min \left(\frac{t^2}{4NA^2}, \frac{t}{2A} \right) \right), \text{ and} \quad (4.4)$$

$$\mathbb{P} \left(\sum_{i \leq N} X_i \geq t \right) \leq \exp \left(-\frac{t^2}{4N\mathbb{E}X^2} \left(1 - \frac{4A^3t}{N(\mathbb{E}X^2)^2} \right) \right). \quad (4.5)$$

It would be interesting to adapt these results to Markov chains, with constants $1/\gamma$ times weaker for empirical averages of reversible chains, and t_{mix} times weaker in general. See [Adamczak \(2008\)](#) for a similar result.

3. (Moment inequalities) [Boucheron et al. \(2005\)](#) proves various moment inequalities for functions of independent random variables using the “entropy method”. It could be interesting to generalize some of these to functions of Markov chains, with constants t_{mix} (or $1/\gamma$) times weaker than in the independent case. We note that [Chazottes and Redig \(2009\)](#) proves moment inequalities for some Markov processes.
4. (Berry-Esseen) Berry-Esseen theorems for empirical averages of discrete Markov chains were proven in [Bolthausen \(1980\)](#), [Mann \(1996\)](#) (for reversible chains, of order $\frac{1/\gamma}{\sqrt{N}}$), and [Lezaud \(1998b\)](#) (for discrete time and continuous time reversible Markov chains, with improved constants compared to [Mann \(1996\)](#), see also [Lezaud \(2001\)](#)).

It would be interesting to get explicit Kolmogorov bounds of order $\sqrt{\frac{1/\gamma}{N}}$ for empirical averages of reversible chains, and $\sqrt{\frac{t_{\text{mix}}}{N}}$ for more general cases.

To convince the reader that this is indeed the correct order, we have the following proposition:

Proposition 4.2. *Let X_1, \dots, X_N be m -dependent, real valued, zero mean random variables, with finite third moments, and let*

$$W = \sum_{i=1}^N X_i, \sigma^2 = \mathbb{E}(W^2).$$

Let $n = \lceil \frac{N}{m} \rceil$, and

$$(Y_1, \dots, Y_{n-1}, Y_n) := (X_1 + \dots + X_m, X_{m+1} + \dots + X_{2m}, \dots, X_{(n-1)m+1} + \dots + X_N).$$

Then we have

$$\sup_z \left| \mathbb{P} \left(\frac{W}{\sigma} \leq z \right) - \Phi(z) \right| \leq \frac{9075}{\sigma^3} \sum_{i=1}^n \mathbb{E}|Y_i|^3 \quad (4.6)$$

Remark 4.1. For m -dependent sequences, we typically have $\sigma \sim \sqrt{mN}$, $\sum_{i=1}^n \mathbb{E}|Y_i|^3 \sim m^2 N$, thus the bound is of order $\sqrt{\frac{m}{N}}$. As far as we know, this is the first bound of this order for m -dependent sequences.

Proof. This is a simple application of Theorem 7.4. of [Barbour and Chen \(2005\)](#) to Y_1, \dots, Y_n . We can also use the same method to show bounds of order $11^{2d} \cdot \sqrt{\frac{m^d}{N}}$ for d dimensional m -dependent random fields. \square

5. (Moderate deviations) Moderate deviations can be seen as an extension of Berry-Esseen to a larger range. Following the notations of Section 11 of [Chen, Goldstein and Shao \(2011\)](#):

Let X_1, \dots, X_n be i.i.d. centered random variables with variance 1, and $W := \frac{X_1 + \dots + X_n}{\sqrt{n}}$. Let $\Phi(z)$ denote the standard normal CDF, then

$$\frac{\mathbb{P}(W \geq z)}{1 - \Phi(z)} = 1 + O(1) \frac{(1 + z^3) \mathbb{E}|X_1|^3}{\sqrt{n}}$$

for $0 \leq z \leq n^{1/6} / (\mathbb{E}|X_1|^3)^{1/3}$.

This result is generalized to several dependence structures using Stein's method (see also [Chen, Fang and Shao \(2009\)](#)).

It would be interesting to prove such a result for functions of Markov chains with explicit constants, since in the range $z \leq n^{1/6} / (\mathbb{E}|X_1|^3)^{1/3}$, it is stronger than the concentration inequalities we got.

6. (DKW) The Dvoretzky-Kiefer-Wolfowitz inequality states the following (this version was proven in [Massart \(1990\)](#)):

Theorem 4.2. Let \hat{F}_N denote the empirical distribution function for a sample of N i.i.d. random variables with distribution function F . Then for every $\lambda > 0$,

$$\mathbb{P} \left(\sqrt{N} \sup_x |\hat{F}_N(x) - F(x)| > \lambda \right) \leq 2 \exp(-2\lambda^2). \quad (4.7)$$

This result means that the Kolmogorov distance of \hat{F}_N and F is typically $O\left(\frac{1}{\sqrt{N}}\right)$, and allows the construction of confidence region for F . It would be interesting to show this result for Markov chains, with constants roughly t_{mix} times weaker than in the independent case ($1/\gamma$ times weaker for reversible chains). A similar result, for geometrically ergodic Markov chains, was proven in [Kontorovich and Weiss \(2012\)](#).

For a simpler exposition of Massart's proof, see [Dudley \(2011\)](#) and Chapter 1 of [Dudley \(2012\)](#).

7. (Bernstein-type DKW) It is clear that for any x sufficiently large, $\hat{F}_N(x)$ has a small variance, and thus by Bernstein's inequality, is closely concentrated around its expectation (in a range much smaller than $\frac{1}{\sqrt{N}}$).

Let $f(X) = \frac{1}{N} \sum_{i \leq N} f_i(X_i)$ with $|f_i(X_i)| \leq C$, then Bernstein's inequality is of the form ($c_1, c_2 > 0, V = \sum_{i \leq N} (f_i(X_i) - \mathbb{E}f_i(X_i))^2$)

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq \exp\left(\frac{-N^2 t^2}{c_1 V + c_2 N C t}\right), \quad (4.8)$$

which is equivalent to

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \geq \frac{c_2 C \tau + \sqrt{(c_2 C \tau)^2 + 4c_1 V \tau / N}}{2N}\right) \leq e^{-\tau}. \quad (4.9)$$

Let $(X_i)_{i \leq N}$ be the random sample, with empirical distribution

$$\hat{F}_N(x) = \frac{1}{N} \sum_{i \leq N} \mathbb{1}[X_i \leq x].$$

Then it is natural to conjecture that for some $c_1, c_2 > 0$,

$$\mathbb{P}\left(|\hat{F}_N(x) - F(x)| \geq \frac{c_2 C \tau + \sqrt{(c_2 C \tau)^2 + 4c_1 V(x) \tau / N}}{2N} \text{ for any } x \in \mathbb{R}\right) \leq e^{-\tau} \quad (4.10)$$

holds, with $V(x) = NF(x)(1-F(x))$ (which could be compared with $\hat{V}(x) = N\hat{F}(x)(1-\hat{F}(x))$ using the original DKW inequality). This would allow sharper confidence regions, especially at the tails.

We think that coupling arguments and breaking the chain into blocks of length t_{mix} (or $1/\gamma$) could be useful. Most of these questions could be directly applied to the analysis of MCMC simulations. We plan to keep up-to-date information about progress on these problems on our webpage.

5. Proofs

The maximal coupling of two random variables is the coupling that achieves the total variational distance of their distribution (see [Lindvall \(1992\)](#), and [Samson \(2000\)](#) page 437):

Definition 8. Let P and Q be two measures defined on a common countable state space Ω . We define the maximal coupling of P and Q , denoted by $\mu_{\max}^{P,Q}$, as

$$\mu_{\max}^{P,Q}(x, y) = \mathbb{1}[x = y] \cdot \min(P(x), Q(y)) + \frac{(P(x) - Q(x))_+ (Q(y) - P(y))_+}{d_{TV}(P, Q)} \quad (5.1)$$

Remark 5.1. The coupling easily generalizes to the non-discrete case. One can see that

$$\begin{aligned} d_{TV}(P, Q) &= \sum_{x \in \Omega} (P(x) - Q(x))_+ = \sum_{y \in \Omega} (Q(y) - P(y))_+ \\ &= 1 - \sum_{x \in \Omega} \min(P(x), Q(x)). \end{aligned} \quad (5.2)$$

Also note that

$$(P(x) - Q(x))_+ (Q(y) - P(y))_+ = (P(x) - Q(x))_+ (Q(y) - P(y))_+ \mathbb{1}[x \neq y]. \quad (5.3)$$

The following lemma will be useful in the proofs:

Lemma 5.1. *Let X, \hat{X}, \mathcal{M} and Γ be as in Theorem 2.1. Let $Y \in \Lambda$, $Y \sim Q$, let \hat{Y} be a partition of Y , similarly to \hat{X} (i.e. $\mathcal{I}(\hat{Y}_i) = \mathcal{I}(\hat{X}_i)$ for $i \leq n$), and $\hat{Y} \sim \hat{Q}$. Then*

$$D(\hat{P}||\hat{Q}) = D(P||Q). \quad (5.4)$$

Let $C(c)$ be as in (2.5), then

$$d_c(Q, P) \leq d_{C(c)}(\hat{Q}, \hat{P}). \quad (5.5)$$

Finally, in the case of d_2 distance,

$$d_2(Q, P) \leq \sqrt{s(\hat{X})} \cdot d_2(\hat{Q}, \hat{P}). \quad (5.6)$$

Proof. (5.4) follows from $P(x) = \hat{P}(\hat{x})$ and $Q(x) = \hat{Q}(\hat{x})$.

$$\begin{aligned} & \mathbb{E}_{\pi(\hat{X} \sim \hat{P}, \hat{Y} \sim \hat{Q})} \left(\sum_{i=1}^n \mathbb{1}[\hat{X}_i \neq \hat{Y}_i] \cdot C_i(c) \right) \\ &= \mathbb{E}_{\pi(\hat{X} \sim \hat{P}, \hat{Y} \sim \hat{Q})} \left(\sum_{i=1}^n \left(\max_{j \in \mathcal{I}_i(\hat{X})} \mathbb{1}[\hat{X}_j \neq \hat{Y}_j] \right) \cdot C_i(c) \right) \\ &\geq \mathbb{E}_{\pi(X \sim P, Y \sim Q)} \left(\sum_{j=1}^N \mathbb{1}[X_j \neq Y_j] \cdot c_j \right). \end{aligned}$$

The infimum of the left hand side is greater or equal to the infimum of the right hand side, so (5.5) follows.

For the d_2 distance, for a coupling $\pi(X \sim P, Y \sim Q)$, denote the “corresponding coupling” of \hat{X}, \hat{Y} by $\hat{\pi}(\hat{X} \sim \hat{P}, \hat{Y} \sim \hat{Q})$ (i.e. $\pi(x, y) = \hat{\pi}(\hat{x}, \hat{y})$), and denote the coupling of these 4 variables by Π .

$$\begin{aligned} & \left(\sum_{\hat{y} \in \hat{\Lambda}} \sum_{i=1}^n \hat{\pi}[\hat{X}_i \neq \hat{y}_i | \hat{Y}_i = \hat{y}_i]^2 \cdot \hat{Q}(\hat{y}) \right)^{1/2} \\ &= \left(\sum_{\hat{y} \in \hat{\Lambda}} \sum_{i=1}^n \mathbb{E}_{\hat{\pi}} \left(\mathbb{1}[\hat{X}_i \neq \hat{y}_i] | \hat{Y}_i = \hat{y}_i \right)^2 \cdot \hat{Q}(\hat{y}) \right)^{1/2} \\ &\geq \left(\left(\sum_{\hat{y} \in \hat{\Lambda}} \sum_{i=1}^n \sum_{j \in \mathcal{I}_i(\hat{X})} \frac{1}{s(\hat{X})} E_{\Pi} \left[\mathbb{1}[X_j \neq y_j] | \hat{Y}_i = \hat{y}_i \right]^2 \right) \cdot Q(\hat{y}) \right)^{1/2} \\ &\geq \left(\frac{1}{s(\hat{X})} \left(\sum_{y \in \Lambda} \sum_{j=1}^N E_{\pi} [\mathbb{1}[X_j \neq y_j] | Y_j = y_j]^2 \right) \cdot Q(y) \right)^{1/2}, \end{aligned}$$

so taking infimum, (5.6) follows. \square

The following coupling construction is a generalization of the construction that has appeared in Samson (2000), Proof of Theorem 1.

Before we start, let us review a simple fact about conditional independence: for 3 discrete random variables X, Y, Z , we say that X is conditionally independent of Y given Z if

$$\begin{aligned} \mathbb{P}(X = x, Y = y, Z = z) \\ &= \mathbb{P}(X = x|Z = z)\mathbb{P}(Y = y|Z = z)\mathbb{P}(Z = z) \\ &= \frac{\mathbb{P}(X = x, Z = z)\mathbb{P}(Y = y, Z = z)}{\mathbb{P}(Z = z)}. \end{aligned} \quad (5.7)$$

Construction 5.1. Given two measures P and Q defined on state space Λ , we are going to define a coupling $\Pi^{P,Q}$ of random variables $Y, X^{(1)}, \dots, X^{(N)}$ taking values in $\Lambda \times \Lambda \times \Lambda_{\geq 2}^N \times \dots \times \Lambda_{\geq n}^N$ with $X^{(1)} \sim P$ and $Y \sim Q$.

Step 1.1. Let $(X_1^{(1)}, Y_1) \sim \mu_{\max}^{P_1, Q_1}$, i.e. the maximal coupling of P_1 and Q_1 .

Step 1.2. Given $(X_1^{(1)}, Y_1)$, we define

$$\begin{aligned} (X_2^{(1)}, \dots, X_N^{(1)} | X_1^{(1)}, Y_1) &\sim P_{\geq 2}^N(\cdot | X_1^{(1)}) \text{ and } (X_2^{(2)}, \dots, X_N^{(2)} | X_1^{(1)}, Y_1) \sim P_{\geq 2}^N(\cdot | Y_1) \text{ as} \\ (X_2^{(2)}, \dots, X_N^{(2)}, X_2^{(1)}, \dots, X_N^{(1)} | Y_1, X_1^{(1)}) &\sim \mathcal{M}^1(\cdot | Y_1, X_1^{(1)}). \end{aligned}$$

In the following, we do similar steps iteratively. Assume that we have already defined $X^{(1)}, \dots, X^{(i)}$ and $Y_{\leq i-1}$ for some $1 < i \leq N$, and that this satisfies

$$(X^{(i)} | Y_{\leq i-1}) \sim P_{\geq i}^N(\cdot | Y_{\leq i-1}).$$

Then we do the following steps:

Step i.1. In this step we add Y_i .

We want

$$(X_i^{(i)}, Y_i | Y_{\leq i-1}) \sim \mu_{\max}^{P_i(\cdot | Y_{\leq i-1}), Q_i(\cdot | Y_{\leq i-1})} \quad (5.8)$$

and we want Y_i to be independent of $X^{(1)}, \dots, X^{(i-1)}, X_{\geq i+1}^{n(i)}$ given $Y_{\leq i-1}, X_i^{(i)}$. From (5.7) one can see that both of these are satisfied by the definition

$$\begin{aligned} \Pi^{P,Q}(y_{\leq i}, x^{(1)}, \dots, x^{(i)}) \\ &:= \frac{\Pi^{P,Q}(y_{\leq i-1}, x^{(1)}, \dots, x^{(i)}) \mu_{\max}^{P_i(\cdot | y_{\leq i-1}), Q_i(\cdot | y_{\leq i-1})}(x_i^{(i)}, y_i)}{P_i(x_i^{(i)} | y_{\leq i-1})} \end{aligned} \quad (5.9)$$

Step i.2. Now we introduce $X^{(i+1)}$ (we skip this step for $i = N$). We want

$$(X^{(i+1)}, X_{\geq i+1}^{n(i)} | Y_{\leq i}, X_i^{(i)}) \sim \mathcal{M}^i(\cdot | Y_{\leq i}, X_i^{(i)}), \quad (5.10)$$

and we want $X^{(i+1)}$ to be independent of $X^{(1)}, \dots, X^{(i-1)}$ given $(X^{(i)}, Y_{\leq i})$. Both conditions are achieved by

$$\begin{aligned} \Pi^{P,Q}(y_{\leq i}, x^{(1)}, \dots, x^{(i+1)}) \\ &:= \frac{\Pi^{P,Q}(y_{\leq i}, x^{(1)}, \dots, x^{(i)}) \cdot \mathcal{M}^i(x^{(i+1)}, x_{\geq i+1}^{n(i)} | y_{\leq i}, x_i^{(i)})}{P_{\geq i+1}^N(x_{\geq i+1}^{n(i)} | y_{\leq i-1}, x_i^{(i)})}. \end{aligned}$$

Iterating these steps up to $i = n$ completes the definition of $\Pi^{P,Q}$.

We continue with the well know tensorization property of the relative entropy:

Lemma 5.2 (Lemma 1 of [Samson \(2000\)](#)). *Let P and Q two probability measures defined on Λ , $E_1 := D(Q_1||P_1)$,*

$$E_i := \sum_{y_{\leq i-1} \in \Lambda_{\leq i-1}} D(Q_i(\cdot|y_{\leq i-1})||P_i(\cdot|y_{\leq i-1})) \cdot Q_{\leq i-1}(y_{\leq i-1}) \quad (5.11)$$

for $2 \leq i \leq N$, then

$$D(Q||P) = \sum_{i \leq N} E_i. \quad (5.12)$$

Finally, we will need a property of the d_2 distance in 1 dimensions:

Lemma 5.3 (Lemma 2 of [Samson \(2000\)](#)). *Let Ω be a discrete state space, and R, W two measures on Ω . Define*

$$d_v(R|W) := \left[\sum_{x \in \Omega} \left[1 - \frac{R(x)}{W(x)} \right]_+^2 W(x) \right]^{1/2},$$

then

$$d_v(R|W)^2 + d_v(W|R)^2 \leq 2D(R||W),$$

and thus

$$d_v(R|W)^2 \leq (2D(R||W))^{1/2}, \quad (5.13)$$

and

$$d_v(W|R)^2 \leq (2D(R||W))^{1/2}. \quad (5.14)$$

Now we are ready to start our proofs:

Proof of Theorem 2.1. First, let us suppose that $\hat{X} = X$ and thus $s(\hat{X}) = 1$ and $N = n$. Let $\Pi := \Pi^{P,Q}$, $(Y, X^{(1)}, \dots, X^{(n)}) \sim \Pi$ (see Construction 5.1), and let $X := X^{(1)}$. Then

$$d_C(Q, P) \leq \mathbb{E}_{\Pi^{P,Q}} \left[\sum_{i \leq n} C_i \mathbb{1}[X_i \neq Y_i] \right].$$

First let us deal with the $i = 1$ term. From step 1.1 in the definition of $\Pi^{P,Q}$, and Pinsker's inequality,

$$\mathbb{E}_{\Pi^{P,Q}} [C_1 \mathbb{1}[X_1 \neq Y_1]] \leq C_1 d_{TV}(P_1, Q_1) \leq C_1 \sqrt{\frac{1}{2} E_1}.$$

For $i = 2$, we can write

$$\mathbb{1}[X_2 \neq Y_2] \leq \mathbb{1}[X_2 \neq X_2^{(2)}] + \mathbb{1}[X_2^{(2)} \neq Y_2].$$

Similarly as before,

$$\begin{aligned} \mathbb{E}_{\Pi} [\mathbb{1}[X_2 \neq Y_2] | Y_1] &= d_{TV}(P_2(\cdot|Y_1), Q_2(\cdot|Y_1)) \\ &\leq \sqrt{\frac{1}{2} D(Q_2(\cdot|Y_1) || P_2(\cdot|Y_1))}, \end{aligned}$$

so by concavity of the square root,

$$\mathbb{E}_{\Pi} (\mathbb{1} [X_2 \neq Y_2]) \leq \sqrt{\frac{1}{2}E_2}.$$

By step 1.2, we can see that $(X^{(2)}, X_{\geq 2}^{n(1)} | Y_1, X_1) \sim \mathcal{M}^1(\cdot | Y_1, X_1)$, thus

$$\mathbb{E}_{\Pi^{P,Q}} [\mathbb{1} [X_2 \neq X_2^{(2)}] | Y_1, X_1] = \mathbb{E}_{\mathcal{M}^1} [\mathbb{1} [X_2 \neq X_2'] | Y_1, X_1] \leq \Gamma_{1,2} \mathbb{1} [X_1 \neq Y_1].$$

From our result for $i = 1$, we get

$$\mathbb{E}_{\Pi^{P,Q}} [\mathbb{1} [X_2 \neq X_2^{(2)}]] \leq \Gamma_{1,2} \cdot \sqrt{\frac{1}{2}E_1}.$$

Similarly, for $i = k$, we can write

$$\begin{aligned} \mathbb{1} [X_k \neq Y_k] &\leq \mathbb{1} [X_k \neq X_k^{(2)}] + \dots + \mathbb{1} [X_k^{(k-1)} \neq X_k^{(k)}] + \mathbb{1} [X_k^{(k)} \neq Y_k] \\ \mathbb{E}_{\Pi} [\mathbb{1} [X_k^{(k)} \neq Y_k]] &\leq \sqrt{\frac{1}{2}E_k} \\ \mathbb{E}_{\Pi} [\mathbb{1} [X_k^{(j)} \neq X_k^{(j+1)}]] &\leq \mathbb{E}_{\Pi} [\mathbb{E}_{\mathcal{M}_k} [\mathbb{1} [X_k \neq X_k'] | Y_{\leq j}, X_j^{(j)}]] \\ &\leq \Gamma_{j,k} \mathbb{E}_{\Pi} [\mathbb{1} [Y_j \neq X_j^{(j)}]] \leq \Gamma_{j,k} \sqrt{\frac{1}{2}E_j} \text{ for } 1 \leq j < k \leq n. \end{aligned}$$

By summing up in i , we get

$$\begin{aligned} d_C(Q, P) &\leq \mathbb{E}_{\Pi} \left[\sum_{i \leq n} C_i \mathbb{1} [X_i \neq Y_i] \right] \\ &\leq \sum_{i \leq n} C_i \sum_{j \leq i} \Gamma_{i,j} \sqrt{\frac{1}{2}E_j} = \sum_{i,j \leq n} C_i \Gamma_{i,j} \sqrt{\frac{1}{2}E_j} \leq \|\Gamma \cdot C\| \sqrt{\frac{1}{2}D(Q||P)}. \end{aligned}$$

The general case, (2.4) follows by Lemma 5.1. \square

Proof of Corollary 2.1. This follows from Theorem 2.1 by Proposition 6.1. of [Ledoux \(2001\)](#). See also Problem 12.3 of [Dubhashi and Panconesi \(2009\)](#). \square

Second proof of Corollary 2.1. Here, we are going to give a direct proof of this result based on the martingale approach of [Chazottes et al. \(2007\)](#) (a similar proof is probably possible using the method of [Kontorovich \(2007\)](#)).

We will write $\hat{f}(\hat{X}) := f(X)$, then as a function of \hat{X} , it is $C(c)$ weighted Hamming Lipschitz (for $\hat{x}, \hat{y} \in \hat{\Lambda}$, $\hat{f}(\hat{x}) - \hat{f}(\hat{y}) \leq d_{C(c)}(\hat{x}, \hat{y})$).

Let us define $\mathcal{F}_i = \sigma(\hat{X}_1, \dots, \hat{X}_i)$ for $i \leq n$, and write

$$f(X) - \mathbb{E}f(X) = \hat{f}(\hat{X}) - \mathbb{E}\hat{f}(\hat{X}) = \sum_{i=1}^n V_i(X),$$

with

$$\begin{aligned}
V_i(\hat{X}) &:= \mathbb{E}(\hat{f}(\hat{X})|\mathcal{F}_i) - \mathbb{E}(\hat{f}(\hat{X})|\mathcal{F}_{i-1}) \\
&= \sum_{z_{\geq i+1}^n} \hat{P}_{\geq i+1}^n(z_{\geq i+1}^n|\hat{X}_{\leq i}) \cdot \hat{f}(\hat{X}_{\leq i}, z_{\geq i+1}^n) \\
&\quad - \sum_{z_{\geq i}^n} \hat{P}_{\geq i}^n(z_{\geq i}^n|\hat{X}_{\leq i-1}) \cdot \hat{f}(\hat{X}_{\leq i-1}, z_{\geq i}^n) \\
&= \sum_{z_{\geq i+1}^n} \hat{P}_{\geq i+1}^n(z_{\geq i+1}^n|\hat{X}_{\leq i}) \cdot \hat{f}(\hat{X}_{\leq i}, z_{\geq i+1}^n) \\
&\quad - \sum_{z_i} \hat{P}_i(z_i|\hat{X}_{\leq i-1}) \sum_{z_{\geq i+1}^n} \hat{P}_{\geq i+1}^n(z_{\geq i+1}^n|\hat{X}_{\leq i-1}, z_i) \cdot \hat{f}(\hat{X}_{\leq i-1}, z_{\geq i}^n) \\
&\leq \sup_{a \in \hat{\Lambda}_i} \sum_{z_{\geq i+1}^n} \hat{P}_{\geq i+1}^n(z_{\geq i+1}^n|\hat{X}_{\leq i-1}, a) \hat{f}(\hat{X}_{\leq i-1}, a, z_{\geq i+1}^n) \\
&\quad - \inf_{b \in \hat{\Lambda}_i} \sum_{z_{\geq i+1}^n} \hat{P}_{\geq i+1}^n(z_{\geq i+1}^n|\hat{X}_{\leq i-1}, b) \hat{f}(\hat{X}_{\leq i-1}, b, z_{\geq i+1}^n) \\
&=: M_i(X) - m_i(X),
\end{aligned}$$

here $M_i(X)$ is the supremum, and $m_i(X)$ is the infimum, and we assume that these values are taken at a and b , respectively (one can take the limit in the following arguments if they do not exist).

After this point, [Chazottes et al. \(2007\)](#) defines a coupling,

$$\pi \left(Z_{\geq i+1}^{n(1)} \sim \hat{P}_{\geq i+1}^n(\cdot|\hat{X}_{\leq i-1}, a), Z_{\geq i+1}^{n(2)} \sim \hat{P}_{\geq i+1}^n(\cdot|\hat{X}_{\leq i-1}, b) \right),$$

as the maximal coupling between these two distributions. Although this coupling minimizes expectation of $\mathbb{1}[Z_{\geq i+1}^{n(1)} \neq Z_{\geq i+1}^{n(2)}]$, it is not always the best choice.

We define

$$\begin{aligned}
&\pi \left(Z_{\geq i+1}^{n(1)} \sim \hat{P}_{\geq i+1}^n(\cdot|\hat{X}_{\leq i-1}, a), Z_{\geq i+1}^{n(2)} \sim \hat{P}_{\geq i+1}^n(\cdot|\hat{X}_{\leq i-1}, b) \right) \\
&:= \mathcal{M}^i \left(Z_{\geq i+1}^{n(1)} \sim \hat{P}_{\geq i+1}^n(\cdot|\hat{X}_{\leq i-1}, a), Z_{\geq i+1}^{n(2)} \sim \hat{P}_{\geq i+1}^n(\cdot|\hat{X}_{\leq i-1}, b) \right).
\end{aligned} \tag{5.15}$$

From this coupling, one can see that

$$\begin{aligned}
M_i(Y) - m_i(Y) &= \mathbb{E}_\pi \left(f(X_{\leq i-1}, Z_{\geq i}^{n(1)}) - f(X_{\leq i-1}, Z_{\geq i}^{n(2)}) | X_{\leq i-1} \right) \\
&\leq \mathbb{E}_\pi \left(\sum_{j=i}^n \mathbb{1}[Z_j^{(1)} \neq Z_j^{(2)}] \cdot C_j \middle| X_1, \dots, X_{i-1} \right) \\
&\leq \sum_{j=i}^n \Gamma_{i,j} C_j.
\end{aligned}$$

The following result was proven in [Devroye and Lugosi \(2001\)](#):

Lemma 5.4. Suppose \mathcal{F} is a sigma-field and Z_1, Z_2, V are random variables such that

1. $Z_1 \leq V \leq Z_2$
2. $\mathbb{E}(V|\mathcal{F}) = 0$
3. Z_1 and Z_2 are \mathcal{F} -measurable.

Then for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}(e^{\lambda V}|\mathcal{F}) \leq e^{\lambda^2(Z_2-Z_1)^2/8}. \quad (5.16)$$

Now using Lemma 5.4 with $V = V_i$, $Z_1 = m_i(X) - \mathbb{E}(\hat{f}(\hat{X})|\mathcal{F}_{i-1})$, $Z_2 = M_i(\hat{X}) - \mathbb{E}(\hat{f}(\hat{X})|\mathcal{F}_{i-1})$, and $\mathcal{F} = \mathcal{F}_{i-1}$, we get that

$$\mathbb{E}\left(e^{\lambda V_i(\hat{X})} \middle| \mathcal{F}_{i-1}\right) \leq \exp\left(\frac{\lambda^2}{8} \left(\sum_{j=i}^n \Gamma_{i,j} C_j(c)\right)^2\right).$$

By taking the product of these, we get

$$\mathbb{E}\left(e^{\lambda \hat{f}(\hat{X})}\right) \leq \exp\left(\frac{\lambda^2}{8} \|\Gamma \cdot C(c)\|^2\right), \quad (5.17)$$

and the tail bound follows by Markov's inequality. \square

Proof of Corollary 2.2. The main idea: we divide the index set into mixing time sized parts. We define the following partition of X : let $n = \left\lceil \frac{N}{\tau(\epsilon)} \right\rceil$, and

$$\begin{aligned} \hat{X} &:= (\hat{X}_1, \dots, \hat{X}_n) \\ &:= ((X_1, \dots, X_{\tau(\epsilon)}), (X_{\tau(\epsilon)+1}, \dots, X_{2\tau(\epsilon)}), \dots, (X_{(n-1)\tau(\epsilon)}, \dots, X_N)). \end{aligned}$$

Such a construction has the following important property: $\hat{X}_1, \dots, \hat{X}_n$ is now a Markov chain, with ϵ mixing time $\hat{\tau}(\epsilon) = 2$ (the proof of this is left to the reader as an exercise).

Now we are going to define a Marton coupling \mathcal{M} for \hat{X} , i.e. for $i \leq n$, we need to define

$$\mathcal{M}^i \left[\hat{X}_{\geq i+1}^n \sim \hat{P}_{\geq i+1}^n(\cdot | \hat{x}_{\leq i}), \hat{X}'_{\geq i+1}^n \sim \hat{P}_{\geq i+1}^n(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i) \right].$$

First step: we define $(\hat{X}_{i+2}, \hat{X}'_{i+2})$ as the maximal coupling of

$$\hat{P}_{i+2}(\cdot | \hat{x}_{\leq i}), \hat{P}_{i+2}(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i),$$

then we have

$$\mathcal{M}^i(\hat{X}_{i+2} \neq \hat{X}'_{i+2} | \hat{x}_{\leq i}, \hat{x}'_i) \leq \epsilon.$$

Second step: we define

$$\mathcal{M}^i \left(\hat{X}_{i+1}, \hat{X}'_{i+1} | \hat{x}_{\leq i}, \hat{x}'_i, \hat{X}_{i+2}, \hat{X}'_{i+2} \right)$$

as the maximal coupling of $\hat{P}_{i+1}(\cdot | \hat{x}_{\leq i}, \hat{X}_{i+2}), \hat{P}_{i+1}(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i, \hat{X}'_{i+2})$. Then trivially

$$\mathcal{M}^i(\hat{X}_{i+1} \neq \hat{X}'_{i+1} | \hat{x}_{\leq i}, \hat{x}'_i) \leq 1.$$

Third step: let $\mathcal{M}^i \left(\hat{X}_{i+4}, \hat{X}'_{i+4} | \hat{x}_{\leq i}, \hat{x}'_i, \hat{X}_{i+1}, \hat{X}'_{i+1}, \hat{X}_{i+2}, \hat{X}'_{i+2} \right)$ be defined as the maximal coupling of

$$\hat{P}_{i+4}(\cdot | \hat{x}_{\leq i}, \hat{X}_{i+1}, \hat{X}_{i+2}), \hat{P}_{i+4}(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i, \hat{X}'_{i+1}, \hat{X}'_{i+2}).$$

By the Markov property, we have

$$\hat{P}_{i+4}(\cdot | \hat{x}_{\leq i}, \hat{X}_{i+1}, \hat{X}_{i+2}) = \hat{P}_{i+4}(\cdot | \hat{X}_{i+2}), \text{ and} \quad (5.18)$$

$$\hat{P}_{i+4}(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i, \hat{X}'_{i+1}, \hat{X}'_{i+2}) = \hat{P}_{i+4}(\cdot | \hat{X}'_{i+2}), \quad (5.19)$$

therefore it is easy to see that

$$\mathcal{M}^i(\hat{X}_{i+4} \neq \hat{X}'_{i+4} | \hat{x}_{\leq i}, \hat{x}'_i) \leq \epsilon^2.$$

Fourth step: we define

$$\mathcal{M}^i \left(\hat{X}_{i+3}, \hat{X}'_{i+3} | \hat{x}_{\leq i}, \hat{x}'_i, \hat{X}_{i+1}, \hat{X}'_{i+1}, \hat{X}_{i+2}, \hat{X}'_{i+2}, \hat{X}_{i+4} \neq \hat{X}'_{i+4} \right)$$

as the maximal coupling of

$$\hat{P}_{i+3}(\cdot | \hat{x}_{\leq i}, \hat{X}_{i+1}, \hat{X}_{i+2}, \hat{X}_{i+4}), \hat{P}_{i+3}(\cdot | \hat{x}_{\leq i-1}, \hat{x}'_i, \hat{X}'_{i+1}, \hat{X}'_{i+2}, \hat{X}'_{i+4}).$$

It is a simple exercise to show that

$$\mathcal{M}^i(\hat{X}_{i+3} \neq \hat{X}'_{i+3} | \hat{x}_{\leq i}, \hat{x}'_i) \leq \epsilon.$$

We get M^i by iterating the third and fourth steps (we can iterate them infinitely, so it is not a problem if $n - i$ is odd). From the construction, it is clear that

$$\Gamma = (\Gamma_{i,j})_{i,j \leq n} \leq \begin{pmatrix} 1 & 1 & \epsilon & \epsilon & \epsilon^2 & \epsilon^2 & \dots \\ 0 & 1 & 1 & \epsilon & \epsilon & \epsilon^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \quad (5.20)$$

with the inequality meant in each element of the matrix.

Now, by the simple fact that $\|\Gamma\| \leq \sqrt{\|\Gamma\|_1 \|\Gamma\|_\infty}$, we have $\|\Gamma\| \leq \frac{2}{1-\epsilon}$, so applying Corollary 2.1 and taking infimum in ϵ proves the result. \square

Proof of Corollary 2.3. This is an immediate consequence of Corollary 2.1. \square

Proof of Corollary 2.4. This follows with 4 times worse constant from Corollary 2.2. We get this better constant by applying a trick from the proof of Theorem 1 of Janson (2004). Without loss of generality, let us assume that N is divisible by $\tau(\epsilon)$, then we can write

$$S = \sum_{i=1}^N f(X_i) = \sum_{j=1}^{\tau(\epsilon)} \sum_{i=1}^{N/\tau(\epsilon)} f(X_{(i-1)\tau(\epsilon)+j}) =: \sum_{j=1}^{\tau(\epsilon)} S_j,$$

i.e. we group $X_{\geq 1}^N$ into $\tau(\epsilon)$ parts. With this notation, it is clear that $S = \sum_{j=1}^{\tau(\epsilon)} S_j$.

Now for some $j \leq \tau(\epsilon)$, $X^j := (X_j, X_{\tau(\epsilon)+j}, \dots, X_{N-\tau(\epsilon)+j})$ forms a Markov chain, with ϵ mixing time $\tau^j(\epsilon) = 1$.

Now for such a chain, we can easily see that the Marton coupling \mathcal{M} for X^j can be constructed by defining \mathcal{M}^i as the maximal coupling, and that we have

$$\Gamma = (\Gamma_{i,j})_{i,j \leq N/\tau(\epsilon)} \leq \begin{pmatrix} 1 & \epsilon & \epsilon^2 & \epsilon^3 & \dots \\ 0 & 1 & \epsilon & \epsilon^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \quad (5.21)$$

Define $c := (b_1 - a_1, \dots, b_N - a_N)$, and $C_j = (b_j - a_j, \dots, b_{N-\tau(\epsilon)+j} - a_{N-\tau(\epsilon)+j})$, for $1 \leq j \leq \tau(\epsilon)$. Then by Corollary 2.1, we have for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}(e^{\lambda S_j}) \leq \exp\left(\frac{\lambda^2}{8} \|\Gamma \cdot C_j\|^2\right) \quad (5.22)$$

$$\leq \exp\left(\frac{\lambda^2}{8} \frac{1}{(1-\epsilon)^2} \|C_j\|^2\right). \quad (5.23)$$

Denote, for $1 \leq j \leq \tau(\epsilon)$, $p_j := \frac{\|C_j\|}{\sum_{j=1}^{\tau(\epsilon)} \|C_j\|}$. Then, by Jensen's inequality, we can write

$$\begin{aligned} \mathbb{E}(e^{\lambda S}) &= \mathbb{E}\left(e^{\lambda \sum_{j=1}^{\tau(\epsilon)} S_j}\right) \\ &= \mathbb{E}\left(e^{\lambda \sum_{j=1}^{\tau(\epsilon)} p_j \frac{1}{p_j} S_j}\right) \leq \sum_{j=1}^{\tau(\epsilon)} p_j \mathbb{E}\left(e^{\lambda \frac{1}{p_j} S_j}\right) \\ &\leq \exp\left(\frac{\lambda^2}{8} \frac{1}{(1-\epsilon)^2} \left(\sum_{j=1}^{\tau(\epsilon)} \|C_j\|\right)^2\right) \\ &\leq \exp\left(\frac{\lambda^2}{8} \frac{1}{(1-\epsilon)^2} \tau(\epsilon) \sum_{j=1}^{\tau(\epsilon)} \|C_j\|^2\right) = \exp\left(\frac{\lambda^2}{8} \frac{1}{(1-\epsilon)^2} \tau(\epsilon) \|c\|^2\right). \end{aligned} \quad (5.24)$$

From this, by Markov inequality we can deduce that

$$\mathbb{P}(S - \mathbb{E}S \geq t) \leq \exp\left(\frac{-2t^2(1-\epsilon)^2}{\|c\|^2 \tau(\epsilon)}\right), \quad (5.25)$$

the same bound holds for the lower tail. Finally, to show (2.11), we need to rescale S by $N - t_0$, and show that $|\mathbb{E}Z - \mathbb{E}_\pi Z| \leq \frac{(b-a)\eta_{\min}(t_0)}{N-t_0}$, these are left to the reader. \square

Proof of Theorem 2.2. Again, let us suppose that $\hat{X} = X$ and thus $s(\hat{X}) = 1$ and $N = n$.

Let $(Y, X^{(1)}, \dots, X^{(n)}) \sim \Pi^{P,Q}$ (see Construction 5.1). For simplicity, in the following, we will write $\Pi := \Pi^{P,Q}$. Then we have

$$d_2(P, Q) \leq \sup_{\alpha: E_Q(\sum_i \alpha_i^2(Y)) \leq 1} \mathbb{E}_\Pi \left(\sum_{i \leq n} \alpha_i(Y) \mathbb{1}[X_i^{(1)} \neq Y_i] \right), \quad (5.26)$$

$$d_2(Q, P) \leq \sup_{\beta: E_P(\sum_i \beta_i^2(X)) \leq 1} \mathbb{E}_\Pi \left(\sum_{i \leq n} \beta_i(X) \mathbb{1}[X_i^{(1)} \neq Y_i] \right). \quad (5.27)$$

For a fixed $\alpha : \Lambda \rightarrow \mathbb{R}$, let us denote $\Delta_i := \mathbb{E}_Q(\alpha_i^2(X))$ (more precisely $\Delta_i(\alpha)$) and similarly $\tilde{\Delta}_i := \mathbb{E}_P(\beta_i^2(X))$.

Then we can assume that $\sum_{i \leq n} \Delta_i \leq 1$ and $\sum_{i \leq n} \tilde{\Delta}_i \leq 1$, since we only take supremum for such α and β .

Let us fix some $i \leq n$, and bound the corresponding term in (5.26):

$$\begin{aligned} & \mathbb{E}_\Pi \left(\alpha_i(Y) \mathbb{1}[X_i^{(1)} \neq Y_i] \right) \leq \\ & \mathbb{E}_\Pi \left(\alpha_i(Y) \left(\mathbb{1}[Y_i \neq X_i^{(i)}] + \mathbb{1}[X_i^{(1)} \neq X_i^{(2)}] + \dots + \mathbb{1}[X_i^{(i-1)} \neq X_i^{(i)}] \right) \right) \\ & =: A_i + \sum_{j=1}^{i-1} B_i^{(j)} \end{aligned} \tag{5.28}$$

$$\begin{aligned} A_i &= \mathbb{E}_\Pi \left(\alpha_i(Y) \mathbb{1}[Y_i \neq X_i^{(i)}] \right) = \sum_{y, x_i^{(i)}} \alpha_i(y) \cdot \mathbb{1} \left[y_i \neq x_i^{(i)} \right] \cdot \Pi \left(y, x_i^{(i)} \right) \\ &= \sum_{y, x_i^{(i)}} \alpha_i(y) \cdot \Pi \left(y_{\geq i+1} \mid x_i^{(i)}, y_{\leq i} \right) \cdot \Pi \left(x_i^{(i)}, y_i \mid y_{\leq i-1} \right) \cdot \Pi(y_{\leq i-1}) \mathbb{1} \left[y_i \neq x_i^{(i)} \right] \\ &= \sum_{y_{\leq i-1}} \Pi(y_{\leq i-1}) \sum_{y_i} \left(\sum_{y_{\geq i+1}^n} \alpha_i(y) \cdot \Pi(y_{\geq i+1} \mid y_{\leq i}) \right) \left(\sum_{x_i^{(i)}} \Pi \left(x_i^{(i)}, y_i \mid y_{\leq i-1} \right) \mathbb{1} \left[y_i \neq x_i^{(i)} \right] \right). \end{aligned}$$

Let us evaluate the last term. By the definition of Y_i in $\Pi^{P,Q}$, we have $(X_i^{(i)}, Y_i \mid Y_{\leq i-1}) \sim \mu_{\max}^{P_i(\cdot \mid Y_{\leq i-1}), Q_i(\cdot \mid Y_{\leq i-1})}$, thus by the definition of the maximal coupling, we get

$$\begin{aligned} & \sum_{x_i^{(i)}} \Pi \left(x_i^{(i)}, y_i \mid y_{\leq i-1} \right) \mathbb{1} \left[y_i \neq x_i^{(i)} \right] \\ &= \sum_{x_i^{(i)}} \frac{[Q_i(y_i \mid y_{\leq i-1}) - P_i(y_i \mid y_{\leq i-1})]_+ \left[P_i(x_i^{(i)} \mid y_{\leq i-1}) - Q_i(x_i^{(i)} \mid y_{\leq i-1}) \right]_+}{d_{TV}(Q_i(\cdot \mid y_{\leq i-1}), P_i(\cdot \mid y_{\leq i-1}))} \\ &= [Q_i(y_i \mid y_{\leq i-1}) - P_i(y_i \mid y_{\leq i-1})]_+, \end{aligned}$$

substituting back gives

$$\begin{aligned} A_i &= \sum_{y_{\leq i-1}} \Pi(y_{\leq i-1}) \cdot \sum_{y_i} \left(\sum_{y_{\geq i+1}^n} \alpha_i(y) \cdot \Pi(y_{\geq i+1} \mid y_{\leq i}) \right) \\ & \quad \cdot [Q_i(y_i \mid y_{\leq i-1}) - P_i(y_i \mid y_{\leq i-1})]_+ \\ &= \sum_{y_{\leq i-1}} Q(y_{\leq i-1}) \sum_{y_i} \left(\sum_{y_{\geq i+1}^n} \alpha_i(y) \cdot Q(y_{\geq i+1} \mid y_{\leq i}) \right) \\ & \quad \cdot \left[1 - \frac{P_i(y_i \mid y_{\leq i-1})}{Q_i(y_i \mid y_{\leq i-1})} \right]_+ Q_i(y_i \mid y_{\leq i-1}) \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{y_{\leq i-1}} Q(y_{\leq i-1}) \left[\sum_{y_i} \left(\sum_{y_{\geq i+1}^n} \alpha_i(y) \cdot Q(y_{\geq i+1} | y_{\leq i}) \right)^2 Q_i(y_i | y_{\leq i-1}) \right]^{1/2} \\
&\cdot \left[\sum_{y_i} \left(1 - \frac{P_i(y_i | y_{\leq i-1})}{Q_i(y_i | y_{\leq i-1})} \right)_+^2 Q_i(y_i | y_{\leq i-1}) \right]^{1/2} \\
&\leq \mathbb{E}_Q \left[\left[\mathbb{E}_Q[\alpha_i^2(Y) | Y_{\leq i-1}] \right]^{1/2} \cdot \left[\mathbb{E}_Q \left[\left(1 - \frac{P_i(Y_i | Y_{\leq i-1})}{Q_i(Y_i | Y_{\leq i-1})} \right)_+^2 \middle| Y_{\leq i-1} \right] \right]^{1/2} \right] \\
&\leq [\mathbb{E}_Q[\alpha_i^2(Y)]]^{1/2} [\mathbb{E}_Q[2D(Q_i(\cdot | Y_{\leq i-1}) || P_i(\cdot | Y_{\leq i-1}))]]^{1/2} \leq (\Delta_i)^{1/2} (2E_i)^{1/2},
\end{aligned}$$

in the last steps we have used Lemma 5.3.

Similarly, for $j < i$,

$$\begin{aligned}
B_i^{(j)} &= \mathbb{E}_\Pi \left(\alpha_i(Y) \mathbb{1} \left[X_i^{(j)} \neq X_i^{(j+1)} \right] \right) \\
&= \sum_{y, x^{(j)}, x^{(j+1)}} \alpha_j(y) \mathbb{1} [x_i^{(j)} \neq x_i^{(j+1)}] \cdot \Pi(y, x^{(j)}, x^{(j+1)}) \\
&= \sum_{y, x^{(j)}, x^{(j+1)}} \alpha_j(y) \mathbb{1} [x_i^{(j)} \neq x_i^{(j+1)}] \cdot \Pi(y_{\leq j-1}) \Pi(x_j^{(j)}, y_j | y_{\leq j-1}) \\
&\cdot \Pi(x_{\geq j+1}^{(j)}, x^{(j+1)} | x_j^{(j)}, y_{\leq j}) \Pi(y_{\geq j+1}^n | y_{\leq j}, x^{(j)}, x^{(j+1)}) \\
&= \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{y_j, x_j^{(j)}} \Pi[x_j^{(j)}, y_j | y_{\leq j-1}] \sum_{x_{\geq j+1}^{(j)}, x^{(j+1)}} \\
&\left[\sum_{y_{\geq j+1}^n} \alpha_i(y) \cdot \Pi(y_{\geq j+1}^n | y_{\leq j}, x^{(j)}, x^{(j+1)}) \right] \cdot \mathbb{1} [x_i^{(j)} \neq x_i^{(j+1)}] \\
&\cdot \Pi \left(x_{\geq j+1}^{(j)}, x^{(j+1)} | x_j^{(j)}, y_{\leq j} \right) \\
&= \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{y_j, x_j^{(j)}} \Pi \left(x_j^{(j)}, y_j | y_{\leq j-1} \right) \cdot \sum_{x_{\geq j+1}^{(j)}, x^{(j+1)}} \mathbb{E}_\Pi(\alpha_i(Y) | y_{\leq j}, x^{(j)}, x^{(j+1)}) \\
&\cdot \mathbb{1} [x_i^{(j+1)} \neq x_i^{(j)}] \Pi \left(x_{\geq j+1}^{(j)}, x^{(j+1)} | x_j^{(j)}, y_{\leq j} \right)
\end{aligned}$$

Applying Cauchy-Schwartz to the last sum gives

$$\begin{aligned}
& \sum_{x_{\geq j+1}^{(j)}, x^{(j+1)}} \mathbb{E}_{\Pi}(\alpha_i(Y)|y_{\leq j}, x^{(j)}, x^{(j+1)}) \cdot \mathbb{1}[x_i^{(j+1)} \neq x_i^{(j)}] \Pi(x_{\geq j+1}^{(j)}, x^{(j+1)}|x_j^{(j)}, y_{\leq j}) \\
& \leq \left(\sum_{x_{\geq j+1}^{(j)}, x^{(j+1)}} \mathbb{E}_{\Pi}(\alpha_i(Y)|y_{\leq j}, x^{(j)}, x^{(j+1)})^2 \cdot \Pi(x_{\geq j+1}^{(j)}, x^{(j+1)}|x_j^{(j)}, y_{\leq j}) \right)^{1/2} \\
& \quad \cdot \left(\sum_{x_{\geq j+1}^{(j)}, x^{(j+1)}} \mathbb{1}[x_i^{(j+1)} \neq x_i^{(j)}] \mathcal{M}^j(x^{(j+1)}, x_{\geq j+1}^{(j)}|y_{\leq j}, x_j^{(j)}) \right)^{1/2} \\
& \leq \left(\mathbb{E}_{\Pi}(\alpha_i(Y)^2 | x_j^{(j)}, y_{\leq j}) \right)^{1/2} \cdot \left(\Gamma_{j,i} \mathbb{1}[x_j^{(j)} \neq y_j] \right)^{1/2}.
\end{aligned}$$

Now we will need the following lemma:

Lemma 5.5. *For any $1 \leq j \leq n-1$, we have*

$$\Pi(y_{\geq j+1}|y_j, x_j^{(j)}) = \Pi(y_{\geq j+1}|y_j).$$

Proof. First, we want to show that $\Pi(y_{j+1}|y_j, x_j^{(j)}) = \Pi(y_{j+1}|y_j)$:

$$\Pi(y_{j+1}|y_{\leq j}) = \sum_{x_{j+1}^{(j+1)}} \Pi(y_{j+1}|y_{\leq j}, x_{j+1}^{(j+1)}) \cdot \Pi(x_{j+1}^{(j+1)}|y_{\leq j})$$

Now by step $(j+1).1$, we have

$$\Pi(y_{j+1}|y_{\leq j}, x_{j+1}^{(j+1)}) = \Pi(y_{j+1}|y_{\leq j}, x_{j+1}^{(j+1)}, x_j^{(j)}),$$

and by step $j.2$, we have

$$\Pi(x_{j+1}^{(j+1)}|y_{\leq j}) = P_{j+1}(x_{j+1}^{(j+1)}|y_{\leq j}) = \Pi(x_{j+1}^{(j+1)}|y_{\leq j}, x_j^{(j)}),$$

thus

$$\Pi(y_{j+1}|y_{\leq j}) = \sum_{x_{j+1}^{(j+1)}} \Pi(y_{j+1}|y_{\leq j}, x_{j+1}^{(j+1)}, x_j^{(j)}) \cdot \Pi(x_{j+1}^{(j+1)}|y_{\leq j}, x_j^{(j)}) = \Pi(y_{j+1}|y_{\leq j}, x_j^{(j)}).$$

The next step is to show that $\Pi(y_{j+2}|y_{\leq j+1}) = \Pi(y_{j+2}|y_{\leq j+1}, x_j^{(j)})$:

$$\begin{aligned}
\Pi(y_{j+2}|y_{\leq j+1}) &= \sum_{x_{j+2}^{(j+2)}} \Pi(y_{j+2}|y_{\leq j+1}, x_{j+2}^{(j+2)}) \cdot \Pi(x_{j+2}^{(j+2)}|y_{\leq j+1}) \\
&= \sum_{x_{j+2}^{(j+2)}} \Pi(y_{j+2}|y_{\leq j+1}, x_{j+2}^{(j+2)}, x_j^{(j)}) \cdot \Pi(x_{j+2}^{(j+2)}|y_{\leq j+1}, x_{j+1}^{(j+1)}, x_j^{(j)})
\end{aligned}$$

$$\begin{aligned}
&= \sum_{x_{j+2}^{(j+2)}} \Pi(y_{j+2}|y_{\leq j+1}, x_{j+2}^{(j+2)}, x_j^{(j)}) \cdot P_{j+2}(x_{j+2}^{(j+2)}|y_{\leq j+1}) \\
&= \sum_{x_{j+2}^{(j+2)}} \Pi(y_{j+2}|y_{\leq j+1}, x_{j+2}^{(j+2)}, x_j^{(j)}) \cdot \Pi(x_{j+2}^{(j+2)}|y_{\leq j+1}, x_j^{(j)}) \\
&= \Pi(y_{j+2}|y_{\leq j+1}, x_j^{(j)}).
\end{aligned}$$

□

Using this lemma, we can now write

$$\begin{aligned}
B_i^{(j)} &\leq \Gamma_{j,i}^{1,2} \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{y_j, x_j^{(j)}} \Pi\left(x_j^{(j)}, y_j \middle| y_{\leq j-1}\right) \cdot \mathbb{1}[x_j^{(j)} \neq y_j] \\
&\quad \cdot (\mathbb{E}_{\Pi}(\alpha_i(Y)^2 | y_{\leq j}))^{1/2}
\end{aligned}$$

By step $j.1$, we can write

$$\begin{aligned}
&\mathbb{1}[x_j^{(j)} \neq y_j] \Pi\left(x_j^{(j)}, y_j \middle| y_{\leq j-1}\right) = \mathbb{1}[x_j^{(j)} \neq y_j] \mu_{\max}^{P_j(\cdot|y_{\leq j-1}), Q_j(\cdot|y_{\leq j-1})} \\
&= \frac{[Q_j(y_j|y_{\leq j-1}) - P_j(y_j|y_{\leq j-1})]_+ \left[P_j(x_j^{(j)}|y_{\leq j-1}) - Q_j(x_j^{(j)}|y_{\leq j-1}) \right]_+}{d_{TV}(Q_j(\cdot|y_{\leq j-1}) - P_j(\cdot|y_{\leq j-1}))},
\end{aligned}$$

thus summing up in $x_j^{(j)}$, we get

$$\begin{aligned}
B_i^{(j)} &\leq \gamma_{j,i} \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{y_j} [Q_j(y_j|y_{\leq j-1}) - P_j(y_j|y_{\leq j-1})]_+ \cdot (\mathbb{E}_{\Pi}(\alpha_i(Y)^2 | y_{\leq j}))^{1/2} \\
&= \gamma_{j,i} \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{y_j} \left[1 - \frac{P_j(y_j|y_{\leq j-1})}{Q_j(y_j|y_{\leq j-1})} \right]_+ \cdot (\mathbb{E}_{\Pi}(\alpha_i(Y)^2 | y_{\leq j}))^{1/2} Q_j(y_j|y_{\leq j-1}) \\
&\leq \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) (\mathbb{E}_{\Pi}(\alpha_i(Y)^2 | y_{\leq j-1}))^{1/2} \cdot \left(\sum_{y_j} \left[1 - \frac{P_j(y_j|y_{\leq j-1})}{Q_j(y_j|y_{\leq j-1})} \right]_+^2 Q_j(y_j|y_{\leq j-1}) \right)^{1/2} \\
&\leq \gamma_{j,i} (\Delta_i)^{1/2} (2E_j)^{1/2}.
\end{aligned}$$

Summing up (5.28) in i gives that

$$\begin{aligned}
&\sum_{i \leq n} \left(A_i + \sum_{j \leq i-1} B_i^{(j)} \right) \leq \sum_{i \leq n} \left((\Delta_i)^{1/2} (2E_i)^{1/2} + \sum_{j \leq i-1} \gamma_{j,i} (\Delta_i)^{1/2} (2E_j)^{1/2} \right) \\
&= \sum_{i,j \leq n} \gamma_{j,i} (\Delta_i)^{1/2} (2E_j)^{1/2} \leq \|\gamma\| \sqrt{2D(Q||P)},
\end{aligned}$$

and since this holds uniformly for all α satisfying

$$E_Q \left(\sum_{i \leq n} \alpha_i^2(Y) \right) = \sum_{i \leq n} \Delta_i \leq 1,$$

(2.14) is proven.

The proof of (2.15) is analogous, we start from (5.27), and continue along the same lines, with β replacing α . Equation (5.28) becomes

$$\begin{aligned} \mathbb{E}_\Pi \left(\beta_i(X) \mathbb{1}[X_i^{(1)} \neq Y_i] \right) &\leq \\ \mathbb{E}_\Pi \left(\beta_i(X) \left(\mathbb{1}[Y_i \neq X_i^{(i)}] + \mathbb{1}[X_i^{(1)} \neq X_i^{(2)}] + \dots + \mathbb{1}[X_i^{(i-1)} \neq X_i^{(i)}] \right) \right) & \\ =: \tilde{A}_i + \sum_{j=1}^{i-1} \tilde{B}_i^{(j)}. \end{aligned} \quad (5.29)$$

We start with the first term:

$$\begin{aligned} \tilde{A}_i &= \mathbb{E}_\Pi \left(\beta_i(X^{(1)}) \mathbb{1}[Y_i \neq X_i^{(i)}] \right) = \sum_{x^{(1)}, y_i, x_i^{(i)}} \beta_i(x^{(1)}) \cdot \mathbb{1}[y_i \neq x_i^{(i)}] \cdot \Pi(x^{(1)}, x_i^{(i)}, y_i) \\ &= \sum_{x^{(1)}, \dots, x^{(i)}, y_{\leq i}} \beta_i(x^{(1)}) \cdot \mathbb{1}[y_i \neq x_i^{(i)}] \cdot \Pi(x^{(1)}, \dots, x^{(i)}, y_{\leq i}) \\ &= \sum_{x^{(1)}, \dots, x^{(i)}, y_{\leq i}} \beta_i(x^{(1)}) \cdot \mathbb{1}[y_i \neq x_i^{(i)}] \cdot \Pi(x^{(1)}, \dots, x^{(i-1)}, x_{\geq i+1}^{n(i)} | x_i^{(i)}, y_{\leq i}) \\ &\quad \cdot \Pi(x_i^{(i)}, y_i | y_{\leq i-1}) \cdot \Pi(y_{\leq i-1}) \end{aligned}$$

Now, from the definition of Π , we can see that

$$\Pi(x^{(1)}, \dots, x^{(i-1)}, x_{\geq i+1}^{n(i)} | x_i^{(i)}, y_{\leq i}) = \Pi(x^{(1)}, \dots, x^{(i-1)}, x_{\geq i+1}^{n(i)} | x_i^{(i)}, y_{\leq i-1}),$$

thus

$$\begin{aligned} \tilde{A}_i &= \sum_{y_{\leq i-1}} \Pi(y_{\leq i-1}) \sum_{x_i^{(i)}, y_i} \left[\sum_{x^{(1)}, \dots, x_{\geq i+1}^{n(i)}} \beta_i(x^{(1)}) \cdot \Pi(x^{(1)}, \dots, x^{(i-1)}, x_{\geq i+1}^{n(i)} | x_i^{(i)}, y_{\leq i-1}) \right] \\ &\quad \cdot \mathbb{1}[y_i \neq x_i^{(i)}] \cdot \Pi(x_i^{(i)}, y_i | y_{\leq i-1}). \end{aligned}$$

We have

$$\begin{aligned} &\mathbb{1}[y_i \neq x_i^{(i)}] \cdot \Pi(x_i^{(i)}, y_i | y_{\leq i-1}) \\ &= \frac{[Q_i(y_i | y_{\leq i-1}) - P_i(y_i | y_{\leq i-1})]_+ \left[P_i(x_i^{(i)} | y_{\leq i-1}) - Q_i(x_i^{(i)} | y_{\leq i-1}) \right]_+}{d_{TV}(Q_i(\cdot | y_{\leq i-1}) - P_i(\cdot | y_{\leq i-1}))}, \end{aligned}$$

summing up in y_i , and then applying Cauchy-Schwartz gives

$$\begin{aligned}
\tilde{A}_i &= \sum_{y_{\leq i-1}} \Pi(y_{\leq i-1}) \sum_{x_i^{(i)}} \left[\sum_{x^{(1)}, \dots, x_{\geq i+1}^{n(i)}} \beta_i(x^{(1)}) \cdot \Pi(x^{(1)}, \dots, x^{(i-1)}, x_{\geq i+1}^{n(i)} | x_i^{(i)}, y_{\leq i-1}) \right] \\
&\quad \cdot \left[P_i(x_i^{(i)} | y_{\leq i-1}) - Q_i(x_i^{(i)} | y_{\leq i-1}) \right]_+ \\
&= \sum_{y_{\leq i-1}} \Pi(y_{\leq i-1}) \sum_{x_i^{(i)}} \left[\sum_{x^{(1)}, \dots, x_{\geq i+1}^{n(i)}} \beta_i(x^{(1)}) \cdot \Pi(x^{(1)}, \dots, x^{(i-1)}, x_{\geq i+1}^{n(i)} | x_i^{(i)}, y_{\leq i-1}) \right] \\
&\quad \cdot \left[1 - \frac{Q_i(x_i^{(i)} | y_{\leq i-1})}{P_i(x_i^{(i)} | y_{\leq i-1})} \right]_+ \cdot P_i(x_i^{(i)} | y_{\leq i-1}) \\
&\leq \sum_{y_{\leq i-1}} [\mathbb{E}_\Pi (\beta_i^2(X^{(1)}) | y_{\leq i-1})]^{1/2} \cdot \left[\sum_{x_i^{(i)}} \left[1 - \frac{Q_i(x_i^{(i)} | y_{\leq i-1})}{P_i(x_i^{(i)} | y_{\leq i-1})} \right]_+^2 \cdot P_i(x_i^{(i)} | y_{\leq i-1}) \right]^{1/2} \\
&\quad \cdot \Pi(y_{\leq i-1}) \leq (2E_i)^{1/2} (\tilde{\Delta}_i)^{1/2}.
\end{aligned}$$

Finally, for $j < i$,

$$\begin{aligned}
\tilde{B}_i^{(j)} &= \mathbb{E} \left(\beta_i(X^{(1)}) \cdot \mathbb{1} \left[X_i^{(j)} \neq X_i^{(j+1)} \right] \right) = \\
&= \sum_{y_{\leq j}, x^{(1)}, \dots, x^{(j+1)}} \beta_i(x^{(1)}) \cdot \mathbb{1} [x_i^{(j)} \neq x_i^{(j+1)}] \cdot \Pi(x^{(1)}, \dots, x^{(j+1)}, y_{\leq j}) \\
&= \sum_{y_{\leq j}, x^{(1)}, \dots, x^{(j+1)}} \beta_i(x^{(1)}) \cdot \mathbb{1} [x_i^{(j)} \neq x_i^{(j+1)}] \cdot \Pi(x^{(1)}, \dots, x^{(j-1)} | x^{(j)}, x^{(j+1)}, y_{\leq j}) \\
&\quad \cdot \Pi(x_{\geq j+1}^{n(j)}, x^{(j+1)} | x_j^{(j)}, y_{\leq j}) \cdot \Pi(x_j^{(j)}, y_j | y_{\leq j-1}) \cdot \Pi(y_{\leq j-1}) \\
&= \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{x_j^{(j)}, y_j} \Pi(x_j^{(j)}, y_j | y_{\leq j-1}) \sum_{x_{\geq j+1}^{n(j)}, x^{(j+1)}} \mathbb{1} [x_i^{(j)} \neq x_i^{(j+1)}] \\
&\quad \cdot \Pi(x_{\geq j+1}^{n(j)}, x^{(j+1)} | x_j^{(j)}, y_{\leq j}) \cdot \sum_{x^{(1)}, \dots, x^{(j-1)}} \beta_i(x^{(1)}) \pi(x^{(1)}, \dots, x^{(j-1)} | x^{(j)}, x^{(j+1)}, y_{\leq j}) \\
&= \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{x_j^{(j)}, y_j} \Pi(x_j^{(j)}, y_j | y_{\leq j-1}) \sum_{x_{\geq j+1}^{n(j)}, x^{(j+1)}} \mathbb{1} [x_i^{(j)} \neq x_i^{(j+1)}] \\
&\quad \cdot \Pi(x_{\geq j+1}^{n(j)}, x^{(j+1)} | x_j^{(j)}, y_{\leq j}) \cdot \mathbb{E}_\Pi (\beta_i(X^{(1)}) | x^{(j)}, x^{(j+1)}, y_{\leq j}) \\
&\leq \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{x_j^{(j)}, y_j} \Pi(x_j^{(j)}, y_j | y_{\leq j-1}) \left(\mathbb{E}_\Pi (\beta_i^2(X^{(1)}) | x_j^{(j)}, y_{\leq j}) \right)^{1/2} \\
&\quad \cdot \left(\mathbb{E}_\Pi \left(\mathbb{1} [X_i^{(j)} \neq X_i^{(j+1)}] \middle| x_j^{(j)}, y_{\leq j} \right) \right)^{1/2}
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{x_j^{(j)}, y_j} \Pi\left(x_j^{(j)}, y_j | y_{\leq j-1}\right) (\Gamma_{j,i})^{1/2} \mathbb{1}[x_j^{(j)} \neq y_j] \\
&\cdot \left(\mathbb{E}_{\Pi}\left(\beta_i^2(X^{(1)}) | x_j^{(j)}, y_{\leq j}\right)\right)^{1/2}.
\end{aligned}$$

Here

$$\mathbb{E}_{\Pi}\left(\beta_i^2(X^{(1)}) | x_j^{(j)}, y_{\leq j}\right) = \mathbb{E}_{\Pi}\left(\beta_i^2(X^{(1)}) | x_j^{(j)}, y_{\leq j-1}\right),$$

as we can see from (5.9), therefore

$$\begin{aligned}
\tilde{B}_i^{(j)} &\leq \gamma_{j,i} \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \\
&\cdot \sum_{x_j^{(j)}, y_j} \Pi\left(x_j^{(j)}, y_j | y_{\leq j-1}\right) \mathbb{1}[x_j^{(j)} \neq y_j] \left(\mathbb{E}_{\Pi}\left(\beta_i^2(X^{(1)}) | x_j^{(j)}, y_{\leq j-1}\right)\right)^{1/2}.
\end{aligned}$$

As previously, we can write

$$\begin{aligned}
&\mathbb{1}[x_j^{(j)} \neq y_j] \Pi\left(x_j^{(j)}, y_j | y_{\leq j-1}\right) = \\
&\frac{[Q_j(y_j | y_{\leq j-1}) - P_j(y_j | y_{\leq j-1})]_+ \left[P_j(x_j^{(j)} | y_{\leq j-1}) - Q_j(x_j^{(j)} | y_{\leq j-1})\right]_+}{d_{TV}(Q_j(\cdot | y_{\leq j-1}) - P_j(\cdot | y_{\leq j-1}))},
\end{aligned}$$

summing up in y_j we get

$$\begin{aligned}
\tilde{B}_i^{(j)} &\leq \gamma_{j,i} \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{x_j^{(j)}} \left[P_j(x_j^{(j)} | y_{\leq j-1}) - Q_j(x_j^{(j)} | y_{\leq j-1})\right]_+ \\
&\cdot \left(\mathbb{E}_{\Pi}\left(\beta_i^2(X^{(1)}) | x_j^{(j)}, y_{\leq j-1}\right)\right)^{1/2} \\
&\leq \gamma_{j,i} \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \sum_{x_j^{(j)}} \left[1 - \frac{Q_j(x_j^{(j)} | y_{\leq j-1})}{P_j(x_j^{(j)} | y_{\leq j-1})}\right]_+ \\
&\cdot \left(\mathbb{E}_{\Pi}\left(\beta_i^2(X^{(1)}) | x_j^{(j)}, y_{\leq j-1}\right)\right)^{1/2} \cdot P_j(x_j^{(j)} | y_{\leq j-1}) \\
&\leq \gamma_{j,i} \sum_{y_{\leq j-1}} \Pi(y_{\leq j-1}) \left(\mathbb{E}_{\Pi}\left(\beta_i^2(X^{(1)}) | y_{\leq j-1}\right)\right)^{1/2} \\
&\cdot \left(\sum_{x_j^{(j)}} \left[1 - \frac{Q_j(x_j^{(j)} | y_{\leq j-1})}{P_j(x_j^{(j)} | y_{\leq j-1})}\right]_+^2 P_j(x_j^{(j)} | y_{\leq j-1})\right)^{1/2} \\
&\leq \gamma_{j,i} (\tilde{\Delta}_i)^{1/2} (2E_j)^{1/2}.
\end{aligned}$$

Summing up in i, j gives (2.15).

The general case, when $\hat{X} \neq X$ and thus $s(\hat{X}) > 1$, follows by Lemma 5.1. \square

Proof of Corollary 2.5. The following proof is based on page 128 of [Ledoux \(2001\)](#) (see also [Dubhashi and Panconesi \(2009\)](#), Section 13.5). First, by the triangle inequality for d_2 distance (see Lemma A and B of [Marton \(2003\)](#)), we have for any distributions Q, R on Λ , $d_2(Q, R) \leq d_2(Q, P) + d_2(P, R)$. Thus, by (2.14) and (2.15), we get

$$\frac{1}{4\|\gamma\|^2 s(\hat{X})} d_2(Q, R)^2 \leq D(R||P) + D(Q||P). \quad (5.30)$$

Now take any $A \subset \Lambda$ with $P(A) > 0$. Let $Q(x) := P(x)/P(A)$ on $x \in A$ and 0 otherwise, then

$$D(Q||P) = \log \left(\frac{1}{P(A)} \right). \quad (5.31)$$

By the definition of Talagrand's convex distance and the d_2 distance, one can see that for any distribution R on Λ , and any Q supported on A ,

$$\mathbb{E}_{Y \sim R} (d_T^2(Y, A)) \leq d_2(Q, R)^2. \quad (5.32)$$

Let

$$Z := \mathbb{E}_P \exp \left(\frac{d_T^2(X, A)}{4\|\gamma\|^2 s(\hat{X})} \right),$$

and

$$R(x) := \frac{1}{Z} \exp \left(\frac{d_T^2(x, A)}{4\|\gamma\|^2 s(\hat{X})} \right) P(x).$$

Using (5.32), we get

$$\begin{aligned} D(R||P) &= \log \left(\frac{1}{Z} \right) + \mathbb{E}_{Y \sim R} \left(\frac{d_T^2(Y, A)}{4\|\gamma\|^2 s(\hat{X})} \right) \\ &\leq \frac{d_2(Q, R)^2}{4\|\gamma\|^2 s(\hat{X})} - \log Z. \end{aligned}$$

Comparing this with (5.30) and (5.31), we get $\log Z \leq \log \left(\frac{1}{P(A)} \right)$, and thus (2.17). \square

Proof of Theorem 2.3. One could prove this result, with constants 4 times worse, by Theorem 2.6. The following proof is similar to the proof of Theorem 2 of [Samson \(2000\)](#).

First, suppose that $X = \hat{X}$, then $s(\hat{X}) = 1$, and $N = n$. We have

$$f(y) - f(x) \leq \sum_{i=1}^N \alpha_i(y) \mathbb{1}[x_i \neq y_i].$$

As previously, denote the law of X by P , and let Y be a random variable taking values in Λ with law Q . Let $\pi[X \sim P, Y \sim Q]$ be a coupling of P and Q . With the shorthand notation $\mathbb{E}_P f := \mathbb{E}_P f(X)$ and $\mathbb{E}_Q f := \mathbb{E}_Q f(X)$, we have

$$\mathbb{E}_Q f - \mathbb{E}_P f = \mathbb{E}_\pi [f(X) - f(Y)] \leq \sum_{x, y \in \Lambda} \sum_{i=1}^N \alpha_i(y) \mathbb{1}[x_i \neq y_i] \pi(x, y)$$

Now using the second definition of $d_2(P, Q)$, (2.13), we can see that

$$\begin{aligned}\mathbb{E}_Q f - \mathbb{E}_P f &\leq \left[E_Q \left(\sum_{i=1}^N \alpha_i^2(Y) \right) \right]^{1/2} \cdot d_2(P, Q), \text{ and similarly,} \\ \mathbb{E}_P f - \mathbb{E}_Q f &\leq \left[E_P \left(\sum_{i=1}^N \alpha_i^2(X) \right) \right]^{1/2} \cdot d_2(Q, P).\end{aligned}$$

By Theorem 2.2, $d_2(P, Q), d_2(Q, P) \leq \|\gamma\| \sqrt{2D(Q||P)}$. Denote $\alpha^2(x) := \sum_{i=1}^N \alpha_i^2(x)$, then

$$\mathbb{E}_Q f - \mathbb{E}_P f \leq [E_Q \alpha^2]^{1/2} \cdot \sqrt{2\|\gamma\|^2 D(Q||P)}, \quad (5.33)$$

$$\mathbb{E}_P f - \mathbb{E}_Q f \leq [E_P \alpha^2]^{1/2} \cdot \sqrt{2\|\gamma\|^2 D(Q||P)}. \quad (5.34)$$

We will now use the following simple lemma, which follows by the Cauchy-Schwartz inequality:

Lemma 5.6. *For $A, B \geq 0$, $\lambda > 0$, we have*

$$\sqrt{AB} \leq \frac{\lambda A}{2} + \frac{B}{2\lambda}.$$

Therefore, we can write, for any $\lambda > 0$,

$$\begin{aligned}\mathbb{E}_Q f - \mathbb{E}_P f &\leq \frac{\lambda \|\gamma\|^2 \mathbb{E}_Q \alpha^2}{2} + \frac{1}{\lambda} D(Q||P), \\ \mathbb{E}_P f - \mathbb{E}_Q f &\leq \frac{\lambda \|\gamma\|^2 \mathbb{E}_P \alpha^2}{2} + \frac{1}{\lambda} D(Q||P).\end{aligned}$$

These can be rewritten as

$$\mathbb{E}_Q \left[\lambda(f - \mathbb{E}_P f) - \frac{\lambda^2 \|\gamma\|^2 \alpha^2}{2} \right] \leq D(Q||P), \quad (5.35)$$

$$\mathbb{E}_Q \left[-\lambda(f - \mathbb{E}_P f) - \frac{\lambda^2 \|\gamma\|^2 \mathbb{E}_P \alpha^2}{2} \right] \leq D(Q||P). \quad (5.36)$$

Now we will need the following lemma:

Lemma 5.7. *Suppose, that a distribution P on Λ satisfies for some function $g : \Lambda \rightarrow \mathbb{R}$ that for every distribution Q on Λ ,*

$$\mathbb{E}_Q g \leq D(Q||P),$$

then $\mathbb{E}_P(e^g) \leq 1$.

Proof. Choose $\frac{Q(x)}{P(x)} = \frac{e^{g(x)}}{\mathbb{E}_P e^{g(X)}}$, then use the definition of $D(Q||P)$. □

Applying this to (5.35) and (5.36), we get

$$\mathbb{E}_P \exp \left[\lambda(f - \mathbb{E}_P f) - \frac{\lambda^2 \|\gamma\|^2 \alpha^2}{2} \right] \leq 1. \quad (5.37)$$

$$\mathbb{E}_P \exp \left[-\lambda(f - \mathbb{E}_P f) - \frac{\lambda^2 \|\gamma\|^2 \mathbb{E}_P \alpha^2}{2} \right] \leq 1. \quad (5.38)$$

Now, using the weakly α -self-bounding condition, we can write $\alpha^2(x) \leq af(x) + b$, thus

$$\mathbb{E}_P \exp \left[\lambda(f - \mathbb{E}_P f - \lambda \|\gamma\|^2 af/2) - \frac{\lambda^2 \|\gamma\|^2 b}{2} \right] \leq 1. \quad (5.39)$$

$$\mathbb{E}_P \exp [-\lambda(f - \mathbb{E}_P f)] \leq \exp \left(\frac{\lambda^2 \|\gamma\|^2 (a\mathbb{E}_P f + b)}{2} \right). \quad (5.40)$$

The tail bounds follow by Markov's inequality (for the first case, with the choice $\lambda = \frac{t}{\|\gamma\|^2 (at + a\mathbb{E}_P f + b)}$).

For later use, for $a > 0$, we will further bound (5.39): we can write

$$\mathbb{E}_P \exp \left[(\lambda - \lambda^2 \|\gamma\|^2 a/2) f \right] \leq \exp \left[\lambda \mathbb{E}_P f + \frac{\lambda^2 \|\gamma\|^2 b}{2} \right]. \quad (5.41)$$

Define the real valued function $d : [0, \frac{1}{2a\|\gamma\|^2}] \rightarrow \mathbb{R}$ as the smallest root of the equation $\lambda - \lambda^2 \|\gamma\|^2 a/2 = z$, i.e.

$$d(z) := \frac{1}{a\|\gamma\|^2} \left(1 - \sqrt{1 - 2a\|\gamma\|^2 z} \right), \quad (5.42)$$

then for $0 \leq z \leq \frac{1}{2a\|\gamma\|^2}$, we have

$$\mathbb{E}_P \exp [zf] \leq \exp \left[d(z) \mathbb{E}_P f + \frac{d(z)^2 \|\gamma\|^2 b}{2} \right]. \quad (5.43)$$

Finally, it is easy to see that for $0 \leq z \leq \frac{1}{2a\|\gamma\|^2}$,

$$d(z) \leq \frac{z}{1 - a\|\gamma\|^2 z}. \quad (5.44)$$

In the general case ($s(\hat{X}) \neq 1$), the right hand side of (5.33) and (5.34) gets multiplied by $\sqrt{s(\hat{X})}$, thus changing from $\|\gamma\|^2$ to $\|\gamma\|^2 s(\hat{X})$ gives the final result. \square

Proof of Corollary 2.6. Notice that if holds, then f is weakly α -(0, C) self-bounding, while if holds, then $-f$ is weakly α -(0, C) self-bounding. Applying Theorem 2.3 proves the result. \square

Proof of Corollary 2.7. The conditions of Corollary 2.6 are satisfied, with $C = 1$. \square

Proof of Corollary 2.8. Let us rewrite (2.26) as

$$Z(x) = \sum_{i \leq N} f_{i, \mathcal{J}(x)}(x_i), \quad (5.45)$$

with

$$\mathcal{J}(x) := \arg \max_{j \leq M} \sum_{i \leq N} f_{i, j}(x_i),$$

then

$$\begin{aligned}
Z(x) - Z(y) &= \sum_{i \leq N} f_{i, \mathcal{J}(x)}(x_i) - \sum_{i \leq N} f_{i, \mathcal{J}(y)}(y_i) \\
&\leq \sum_{i \leq N} f_{i, \mathcal{J}(x)}(x_i) - \sum_{i \leq N} f_{i, \mathcal{J}(x)}(y_i) \\
&\leq \sum_{i \leq N} f_{i, \mathcal{J}(x)}(x_i) \cdot \mathbb{1}[x_i \neq y_i].
\end{aligned}$$

Now $0 \leq f(x) \leq C$, thus the α -(1, 0) self-boundedness of Z/C follows. \square

Proof of Theorem 2.4. First, suppose that $X = \hat{X}$, then $s(\hat{X}) = 1$, and $N = n$. Without loss of generality, suppose that $C = 1$ (changing to $Z(x)/C$ solves the general case). Then

$$S(y) - S(x) \leq \sum_{i=1}^N [(f_i(y_i))_+ + (f_i(x_i))_-] \cdot \mathbb{1}[x_i \neq y_i], \quad (5.46)$$

thus, similarly to the proof of Theorem 2.3, we can write, for any measure Q on Λ ,

$$\begin{aligned}
\mathbb{E}_Q S - \mathbb{E}_P S &\leq \sqrt{E_Q \left[\sum_{i \leq N} (f_i(Y_i)_+)^2 \right]} d_2(Q, P) + \sqrt{E_P \left[\sum_{i \leq N} (f_i(X_i)_-)^2 \right]} d_2(P, Q) \leq \\
&\left(\sqrt{E_Q V_+} + \sqrt{E_P V_-} \right) \sqrt{2 \|\gamma\|^2 D(Q \| P)}, \quad (5.47)
\end{aligned}$$

where $V_+(x) := \sum_{i \leq N} (f_i(x_i)_+)^2$ and $V_-(x) := \sum_{i \leq N} (f_i(x_i)_-)^2$.

Now by Lemma 5.6, we get

$$\begin{aligned}
\mathbb{E}_Q S - \mathbb{E}_P S &\leq \frac{\lambda \|\gamma\|^2}{2} (E_Q(V_+) + E_P(V_-)) + \frac{2}{\lambda} D(Q \| P), \\
\mathbb{E}_Q \left[\frac{\lambda}{2} (S - \mathbb{E}_P S) - \frac{\lambda^2 \|\gamma\|^2}{4} (V_+ + E_P(V_-)) \right] &\leq D(Q \| P).
\end{aligned}$$

By Lemma (5.7), we get

$$\mathbb{E}_P \exp \left[\frac{\lambda}{2} (S - \mathbb{E}_P S) - \frac{\lambda^2 \|\gamma\|^2}{4} (V_+ + E_P(V_-)) \right] \leq 1.$$

Now we will use a simple lemma:

Lemma 5.8. *Let $A, B > 0$ be random variables defined on the same probability space. If $\mathbb{E}(A/B) \leq 1$, then $\mathbb{E}(A^{1/2}) \leq \mathbb{E}(B^{1/2})$.*

Proof. $A^{1/2} = (A/B)^{1/2} \cdot B^{1/2}$, so applying Cauchy-Schwartz gives the result. \square

By this lemma, we get

$$\mathbb{E}_P \exp \left[\frac{\lambda}{4} (S - \mathbb{E}_P S) \right] \leq \left[\mathbb{E}_P \exp \left(\frac{\lambda^2 \|\gamma\|^2}{4} V_+ \right) \right]^{1/2} \cdot \left[\frac{\lambda^2 \|\gamma\|^2}{8} E_P V_- \right], \text{ so } \quad (5.48)$$

$$\mathbb{E}_P \exp [\lambda (S - \mathbb{E}_P S)] \leq \left[\mathbb{E}_P \exp (4 \lambda^2 \|\gamma\|^2 V_+) \right]^{1/2} \cdot \exp [2 \lambda^2 \|\gamma\|^2 E_P V_-]. \quad (5.49)$$

Now V_- is (1,0) - self-bounding, so, by (5.43) and (5.44), we have for $0 \leq z \leq \frac{1}{2\|\gamma\|^2}$,

$$\mathbb{E}_P \exp [zV_+] \leq \exp \left[\frac{z}{1 - \|\gamma\|^2 z} \mathbb{E}_P V_+ \right].$$

With the choice $z = 4\lambda^2 \|\gamma\|^2$, we have, for $0 \leq \lambda \leq \frac{1}{2\sqrt{2}\|\gamma\|^2}$,

$$\mathbb{E}_P \exp (4\lambda^2 \|\gamma\|^2 V_+) \leq \exp \left[\frac{4\lambda^2 \|\gamma\|^2}{1 - 4\|\gamma\|^4 \lambda^2} \mathbb{E}_P V_+ \right] \leq \exp \left[\frac{4\lambda^2 \|\gamma\|^2}{1 - 2\|\gamma\|^2 \lambda} \mathbb{E}_P V_+ \right].$$

Combining this with $\mathbb{E}(V_-) + \mathbb{E}(V_+) \leq V$, and (5.49), we get, for $0 \leq \lambda \leq \frac{1}{2\sqrt{2}\|\gamma\|^2}$,

$$\mathbb{E}_P \exp [\lambda(S - \mathbb{E}_P S)] \leq \exp \left[\frac{2\|\gamma\|^2 V \lambda^2}{1 - 2\sqrt{2}\|\gamma\|^2 \lambda} \right].$$

We get the tail bounds by the following simple lemma:

Lemma 5.9. *Let $G(\lambda) := \log(\mathbb{E} e^{\lambda(f(X) - \mathbb{E} f(X))})$. If for every $\lambda > 0$,*

$$G(\lambda) \leq \frac{c_1 \lambda^2}{1 - c_2 \lambda}, \quad (5.50)$$

for some $c_1, c_2 \geq 0$, then for every $t > 0$,

$$\mathbb{P}(f(X) - \mathbb{E} f(X) \geq t) \leq \exp \left(\frac{-t^2}{4c_1 + 2c_2 t} \right). \quad (5.51)$$

Proof. Apply Markov's inequality for $\lambda = \frac{t}{2c_1 + c_2 t}$. \square

The general case ($s(\hat{X}) \neq 1$) follows by applying the general version of Theorem 2.2 in (5.47), and changing $\|\gamma\|^2$ to $\|\gamma\|^2 s(\hat{X})$. \square

Proof of Corollary 2.9. This is similar to the proof of Corollary 2.4, using (2.30). \square

Proof of Theorem 2.5. The proof is similar to the proof of Theorem 2.4. A different proof, with slightly worse constants, is possible using Theorem 2.6.

Again, first suppose that $s(\hat{X}) = 1$.

Let us reformulate (2.36) as

$$Z(x) = \sum_{i \leq N} f_{i, \mathcal{J}(x)}(x_i), \quad (5.52)$$

with

$$\mathcal{J}(x) := \arg \max_{j \leq M} \sum_{i \leq N} f_{i, j}(x_i).$$

Now

$$\begin{aligned}
Z(x) - Z(y) &= \sum_{i \leq N} f_{i, \mathcal{I}(y)}(y_i) - \sum_{i \leq N} f_{i, \mathcal{I}(x)}(x_i) \\
&\leq \sum_{i \leq N} f_{i, \mathcal{I}(y)}(y_i) - \sum_{i \leq N} f_{i, \mathcal{I}(y)}(x_i) \\
&\leq \sum_{i \leq N} [f_{i, \mathcal{I}(y)}(y_i) - f_{i, \mathcal{I}(y)}(x_i)] \cdot \mathbb{1}[x_i \neq y_i] \\
&\leq \sum_{i \leq N} \left[(f_{i, \mathcal{I}(y)}(y_i))_+ + \max_{j \leq M} (f_{i, j}(x_i))_- \right] \cdot \mathbb{1}[x_i \neq y_i],
\end{aligned}$$

Define $W_+(x) := \sum_{i \leq N} \max_{j \leq M} (f_{i, j}(x_i))_+^2$, and $W_-(x) := \sum_{i \leq N} \max_{j \leq M} (f_{i, j}(y_i))_-^2$, then

$$\begin{aligned}
\mathbb{E}_Q Z - \mathbb{E}_P Z &\leq \sqrt{E_Q \left[\sum_{i \leq N} ((f_{i, \mathcal{I}(Y)}(Y_i))_+)^2 \right] d_2(Q, P)} \\
&+ \sqrt{E_P \left[\sum_{i \leq N} (\max_{j \leq M} (f_{i, j}(X_i))_-)^2 \right] d_2(P, Q)} \leq \\
&\left(\sqrt{E_Q W_+} + \sqrt{E_P W_-} \right) \sqrt{2 \|\gamma\|^2 D(Q \| P)}.
\end{aligned}$$

From here, similar arguments as in the proof of 2.4, and the fact that $\mathbb{E}(W_+) \leq \mathbb{E}(W_-) \leq W$, lead to

$$\mathbb{E}_P \exp [\lambda(Z - \mathbb{E}_P Z)] \leq \exp \left[\frac{4 \|\gamma\|^2 W \lambda^2}{1 - 2\sqrt{2} \|\gamma\|^2 \lambda} \right],$$

and thus the tail bounds follow by Lemma 5.9. The proof for the lower tail is similar.

The general case ($s(\hat{X}) \neq 1$) follows by replacing $\|\gamma\|^2$ with $\|\gamma\|^2 s(\hat{X})$. The proof for (2.41) is similar, except $W_+(x)$ and $W_-(x)$ are replaced by $\sum_{i \leq N} \max_{j \leq M} (f_{i, j}(x_i))^2$, but this does not change the result (since we have already bounded their expected value by W). \square

Proof of Theorem 2.6. We can obviously define $\hat{f} : \hat{\Lambda} \rightarrow \mathbb{R}$ such that $\hat{f}(\hat{x}) = f(x)$ for every $x \in \Lambda$. Suppose first that f satisfies Condition 1, then \hat{f} satisfies, for every $\hat{x}, \hat{y} \in \hat{\Lambda}$,

$$\hat{f}(\hat{x}) - \hat{f}(\hat{y}) \leq \sum_{i=1}^n \hat{\alpha}_i(\hat{x}) \mathbb{1}[\hat{x}_i \neq \hat{y}_i], \quad (5.53)$$

with $\hat{\alpha}_i(\hat{x}) := \sum_{j \in \mathcal{I}_i(\hat{X})} \alpha_j(x)$.

Similarly, if f satisfies Condition 1, then \hat{f} satisfies, for every $\hat{x}, \hat{y} \in \hat{\Lambda}$,

$$\hat{f}(\hat{x}) - \hat{f}(\hat{y}) \leq \sum_{i=1}^n \left(\hat{\alpha}_i(\hat{x}) + \hat{\beta}_i(\hat{y}) \right) \mathbb{1}[\hat{x}_i \neq \hat{y}_i], \quad (5.54)$$

with $\hat{\alpha}_i(\hat{x}) := \sum_{j \in \mathcal{I}_i(\hat{X})} \alpha_j(x)$ and $\hat{\beta}_i(\hat{x}) := \sum_{j \in \mathcal{I}_i(\hat{X})} \beta_j(x)$.

Then, with these notations, it is clear that

$$\begin{aligned} V_{\hat{\alpha}} &:= \mathbb{E}_{\hat{P}} \sum_{i=1}^n \hat{\alpha}_i^2(X) \leq s(\hat{X})V_{\alpha}, \\ V_{\hat{\beta}} &:= \mathbb{E}_{\hat{P}} \sum_{i=1}^n \hat{\beta}_i^2(X) \leq s(\hat{X})V_{\beta}, \\ g_{\hat{\alpha}}(\tau) &:= \log \mathbb{E}_{\hat{P}} e^{\tau \sum_{i=1}^n \hat{\alpha}_i^2(X)} \leq g_{\alpha}(s(\hat{X})\tau), \\ g_{\hat{\beta}}(\tau) &:= \log \mathbb{E}_{\hat{P}} e^{\tau \sum_{i=1}^n \hat{\beta}_i^2(X)} \leq g_{\beta}(s(\hat{X})\tau). \end{aligned}$$

Therefore, in the following, we can make this assumption without loss of generality:

Assumption 5.1. $\hat{X} = X$, and thus $N = n$ and $s(\hat{X}) = X$.

Define, for $\lambda > 0$, $x \in \Lambda$,

$$\mu_{\lambda}(x) := \frac{\exp(\lambda(f(x) - \mathbb{E}f))}{F(\lambda)} P(x), \quad (5.55)$$

$$\nu_{\lambda}(x) := \frac{\exp(-\lambda(f(x) - \mathbb{E}f))}{F(-\lambda)} P(x). \quad (5.56)$$

Now we divide our argument into two parts, depending on which condition on f holds.

Proof for Condition 1. We will use the following lemma:

Lemma 5.10. *Let $Y \sim \mu_{\lambda}$, then*

$$D(\mu_{\lambda}||P) \leq 2\lambda^2 \|\gamma\|^2 \cdot \mathbb{E}_{\mu_{\lambda}} \sum_{i=1}^n \alpha_i^2(Y). \quad (5.57)$$

Proof. First step:

$$\begin{aligned} D(\mu_{\lambda}||P) &= \sum_{x \in \Lambda} \log \left(\frac{\mu_{\lambda}(x)}{P(x)} \right) \mu_{\lambda}(x) \\ &= \sum_{x \in \Lambda} (\lambda[f(x) - \mathbb{E}f] - \log F(\lambda) + \log P(x) - \log P(x)) \mu_{\lambda}(x) \\ &= \lambda[\mathbb{E}_{\mu_{\lambda}} f(Y) - \mathbb{E}_P f(X)] - \log F(\lambda) \\ &\leq \lambda[\mathbb{E}_{\mu_{\lambda}} (f(Y)) - \mathbb{E}_P (f(X))]. \end{aligned} \quad (5.58)$$

By (5.54), we can further bound this as

$$\begin{aligned} \lambda[\mathbb{E}_{\mu_{\lambda}} (f(Y)) - \mathbb{E}_P (f(X))] &= \lambda[\mathbb{E}_{\pi(X \sim P, Y \sim \mu_{\lambda})} (f(Y) - f(X))] \\ &\leq \lambda \cdot \mathbb{E}_{\pi(X \sim P, Y \sim \mu_{\lambda})} \left(\sum_{i=1}^n \alpha_i(Y) \cdot \mathbb{1}[X_i \neq Y_i] \right). \end{aligned}$$

By (2.13), we have

$$\begin{aligned} d_2(\mu_\lambda, P) &= \inf_{\pi(X \sim P, Y \sim \mu_\lambda)} \sup_{a: \mathbb{E}_{\mu_\lambda} \sum a_i^2(Y) \leq 1} \mathbb{E}_\pi \left(\sum_{i=1}^n a_i(Y) \mathbb{1}[X_i \neq Y_i] \right) \\ &= \inf_{\pi(X \sim P, Y \sim \mu_\lambda)} \sup_a \frac{1}{(\mathbb{E}_{\mu_\lambda} \sum a_i^2(Y))^{1/2}} \mathbb{E}_\pi \left(\sum_{i=1}^n a_i(Y) \mathbb{1}[X_i \neq Y_i] \right), \end{aligned}$$

thus

$$\begin{aligned} D(\mu_\lambda \| P) &\leq \lambda \left(\mathbb{E}_{\mu_\lambda} \sum \alpha_i^2(Y) \right)^{1/2} d_2(\mu_\lambda, P) \\ &\leq \lambda \left(\mathbb{E}_{\mu_\lambda} \sum \alpha_i^2(Y) \right)^{1/2} \|\gamma\| \sqrt{2D(\mu_\lambda \| P)}, \end{aligned} \tag{5.59}$$

the statement follows by rearrangement. \square

We need to further bound $\mathbb{E}_{\mu_\lambda} \sum_{i=1}^n \alpha_i^2(Y)$:

Lemma 5.11. *For any $\tau > 0$,*

$$\mathbb{E}_{\mu_\lambda} \left(\sum_{i=1}^n \alpha_i^2(Y) \right) \leq \frac{1}{\tau} (D(\mu_\lambda \| Q) + g_\alpha(\tau)). \tag{5.60}$$

Proof. Let us define Q as

$$Q(x) := \frac{\exp(\tau \sum_{i=1}^n \alpha_i^2(x))}{\mathbb{E}_P(\exp(\tau \sum_{i=1}^n \alpha_i^2(X)))} \cdot P(x),$$

then

$$\begin{aligned} 0 \leq D(\mu_\lambda \| Q) &= \sum_{x \in \Lambda} \mu_\lambda(x) \log \left(\frac{\mu_\lambda(x)}{Q(x)} \right) = \\ &= \sum_{x \in \Lambda} \mu_\lambda(x) \left(\log(\mu_\lambda(x)) - \log(P(x)) - \tau \sum_{i=1}^n \alpha_i^2(x) + \log \left(\mathbb{E}_P e^{\tau \sum_{i=1}^n \alpha_i^2(X)} \right) \right) \\ &= D(\mu_\lambda \| P) - \tau \mathbb{E}_{\mu_\lambda} \left(\sum_{i=1}^n \alpha_i^2(X) \right) + \log \left(\mathbb{E}_P \left(e^{\tau \sum_{i=1}^n \alpha_i^2(X)} \right) \right), \end{aligned}$$

so we have

$$\tau \mathbb{E}_{\mu_\lambda} \left(\sum_{i=1}^n \alpha_i^2(X) \right) \leq D(\mu_\lambda \| P) + \log \left(\mathbb{E}_P \left(e^{\tau \sum_{i=1}^n \alpha_i^2(X)} \right) \right),$$

and thus (5.60) follows. \square

Combining the two lemmas, we get

$$\begin{aligned} D(\mu_\lambda \| P) &\leq 2\lambda^2 \|\gamma\|^2 \cdot \frac{1}{\tau} (D(\mu_\lambda \| Q) + g_\alpha(\tau)) \\ D(\mu_\lambda \| P) &\leq \frac{2\lambda^2 \|\gamma\|^2 \cdot g_\alpha(\tau)}{\tau - 2\lambda^2 \|\gamma\|^2}, \end{aligned} \tag{5.61}$$

for every $\lambda > 2\lambda^2\|\gamma\|^2$.

We continue with the Herbst argument (we use (5.58)):

$$\begin{aligned} \frac{d}{d\lambda} \left(\frac{G(\lambda)}{\lambda} \right) &= \frac{\lambda F'(\lambda)/F(\lambda) - G(\lambda)}{\lambda^2} \\ &= \frac{1}{\lambda} [\mathbb{E}_{\mu_\lambda} f(Y) - \mathbb{E}_P f(X)] - \frac{1}{\lambda^2} \log F(\lambda) \\ &= \frac{1}{\lambda^2} D(\mu_\lambda \| P) \leq \frac{2\|\gamma\|^2 \cdot g_\alpha(\tau)}{\tau - 2\lambda^2\|\gamma\|^2}. \end{aligned}$$

The right hand side is increasing in λ , and $\lim_{\lambda \rightarrow 0} \frac{G(\lambda)}{\lambda} = 0$, therefore

$$G(\lambda) \leq \frac{2\lambda^2\|\gamma\|^2 \cdot g_\alpha(\tau)}{\tau - 2\lambda^2\|\gamma\|^2}, \quad (5.62)$$

thus (2.42) follows.

For the lower tail, we will need

Lemma 5.12.

$$D(\nu_\lambda \| P) \leq 2\lambda^2\|\gamma\|^2 \cdot \mathbb{E}_P \sum_{i=1}^n \alpha_i^2(X). \quad (5.63)$$

Proof.

$$\begin{aligned} D(\nu_\lambda \| P) &= \sum_{x \in \Lambda} \log \left(\frac{\nu_\lambda(x)}{P(x)} \right) \nu_\lambda(x) \\ &= \sum_{x \in \Lambda} (-\lambda [f(x) - \mathbb{E}f] - \log F(-\lambda) + \log P(x) - \log P(x)) \nu_\lambda(x) \\ &= -\lambda [\mathbb{E}_{\nu_\lambda} f(Y) - \mathbb{E}_P f(X)] - \log F(-\lambda) \\ &\leq -\lambda [\mathbb{E}_{\nu_\lambda} (f(Y)) - \mathbb{E}_P (f(X))]. \end{aligned} \quad (5.64)$$

By (5.54) and (2.13), we can further bound this as

$$\begin{aligned} \lambda [\mathbb{E}_P (f(X)) - \mathbb{E}_{\nu_\lambda} (f(Y))] &= \lambda [\mathbb{E}_{\pi(X \sim P, Y \sim \nu_\lambda)} (f(X) - f(Y))] \\ &\leq \lambda \cdot \mathbb{E}_{\pi(X \sim P, Y \sim \nu_\lambda)} \left(\sum_{i=1}^n \alpha_i(X) \cdot \mathbb{1}[X_i \neq Y_i] \right) \\ &\leq \lambda \left(\mathbb{E}_P \sum \alpha_i^2(X) \right)^{1/2} d_2(P, \nu_\lambda) \\ &\leq \lambda \left(\mathbb{E}_P \sum \alpha_i^2(X) \right)^{1/2} \|\gamma\| \sqrt{2D(\nu_\lambda \| P)}. \end{aligned}$$

□

The Herbst argument in this case (see (5.64)):

$$\begin{aligned} \frac{d}{d\lambda} \left(\frac{G(-\lambda)}{\lambda} \right) &= \frac{-\lambda F'(-\lambda)/F(-\lambda) - G(-\lambda)}{\lambda^2} \\ &= -\frac{1}{\lambda} [\mathbb{E}_{\nu_\lambda} f(Y) - \mathbb{E}_P f(X)] - \frac{1}{\lambda^2} \log F(-\lambda) \\ &= \frac{1}{\lambda^2} D(\nu_\lambda \| P) \leq 2\|\gamma\|^2 V_\alpha, \end{aligned}$$

and thus (2.43) follows by integration. □

Proof for condition 2. The proof is based on the following lemma:

Lemma 5.13. *For every $\lambda > 0, \tau > 4\lambda^2\|\gamma\|^2$,*

$$D(\mu_\lambda||P) \leq \frac{4\lambda^2\|\gamma\|^2}{\tau - 4\lambda^2\|\gamma\|^2} (g_\alpha(\tau) + \tau V_\beta). \quad (5.65)$$

Proof.

$$\begin{aligned} D(\mu_\lambda||P) &\leq \lambda \cdot [\mathbb{E}_{\mu_\lambda} f(Y) - \mathbb{E}_P f(X)] \\ &\leq \lambda \mathbb{E}_{\pi(X \sim P, Y \sim \mu_\lambda)} \left(\sum_{i=1}^n (\alpha_i(Y) + \beta_i(X)) \mathbb{1}[X_i \neq Y_i] \right) \\ &\leq \lambda \|\gamma\| \left(\left[\mathbb{E}_{\mu_\lambda} \left(\sum_{i=1}^n \alpha_i^2(Y) \right) \right]^{1/2} + \left[\mathbb{E}_P \left(\sum_{i=1}^n \beta_i^2(X) \right) \right]^{1/2} \right) \\ &\quad \cdot \sqrt{2D(\mu_\lambda||P)}. \end{aligned}$$

By Lemma 5.11, we get

$$\mathbb{E}_{\mu_\lambda} \left(\sum_{i=1}^n \alpha_i^2(Y) \right) \leq \frac{1}{\tau} (D(\mu_\lambda||Q) + g_\alpha(\tau)),$$

and thus we can write

$$\begin{aligned} D(\mu_\lambda||P) &\leq \lambda \|\gamma\| \left(\left(\frac{1}{\tau} (D(\mu_\lambda||Q) + g_\alpha(\tau)) \right)^{1/2} + \left[\mathbb{E}_P \left(\sum_{i=1}^n \beta_i^2(X) \right) \right]^{1/2} \right) \\ &\quad \cdot \sqrt{2D(\mu_\lambda||P)} \\ D(\mu_\lambda||P) &\leq 4\lambda^2\|\gamma\|^2 \left(\frac{1}{\tau} (D(\mu_\lambda||Q) + g_\alpha(\tau)) + V_\beta \right) \\ D(\mu_\lambda||P)(\tau - 4\lambda^2\|\gamma\|^2) &\leq 4\lambda^2\|\gamma\|^2 (g_\alpha(\tau) + \tau V_\beta), \end{aligned}$$

this implies (5.65). □

Now (2.44) follows by the Herbst argument. The proof of (2.45) is similar. □

Proof of Corollary 2.10. Let X_1^*, \dots, X_n^* be a stationary Markov chain with the same probability transition matrix, then it is easy to see that □

$$d_{TV}(\mathcal{L}(X_{t_0+1}), \mathcal{L}(X_{t_0+1}^*)) \leq \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor},$$

therefore, by the Markov property, we can make a coupling between (X_{t_0+1}, \dots, X_n) and $(X_{t_0+1}^*, \dots, X_n^*)$ such that

$$\mathbb{P}((X_{t_0+1}, \dots, X_n) \neq (X_{t_0+1}^*, \dots, X_n^*)) \leq \inf_{0 \leq \epsilon < 1} \epsilon^{\lfloor \frac{t_0}{t_{\text{mix}}(\epsilon)} \rfloor}.$$

Now applying Theorem 1.1 of [Lezaud \(1998a\)](#) to $X_{t_0+1}^*, \dots, X_n^*$, and using Proposition 1.1 proves the first result.

For the second result, we need to modify the proof of Theorem 1.2 on page 47 of [Lezaud \(1998b\)](#).

Lemma 1.1. of [Lezaud \(1998b\)](#) shows that for any $r > 0$,

$$\mathbb{P}[Z - \mathbb{E}_\pi f \geq t] \leq e^r N_q \exp\{-n(rt - \log(\beta_0(P(r))))\}, \quad (5.66)$$

here $\beta_0(P(r))$ having the following Taylor expansion, for $0 \leq r \leq \frac{\gamma}{3}$:

$$\beta_0(P(r)) = 1 + \sum_{n=1}^{\infty} \beta^{(n)} r^n,$$

with $\beta^{(1)} = 0$, $\beta^{(2)} = \sigma^2/2$, and $\beta^{(n)} \leq (V_f/5)(5/\gamma)^{n-1}$ for $n \geq 3$.

From this point (using Proposition 1.5 on page 48, which shows that $\sigma^2 \leq 2V_f/\gamma$), [Lezaud \(1998b\)](#) shows that $\beta_0(P(r)) \leq 1 + \frac{V_f}{\gamma} r^2 \left(1 - \frac{5r}{\gamma}\right)^{-1}$, and thus obtains a result depending on V_f and γ only. Here, we take a different approach:

$$\beta_0(P(r)) \leq 1 + \frac{\sigma^2}{2} r^2 + \sum_{n=3}^{\infty} (V_f/5)(5/\gamma)^{n-1} r^n \leq 1 + \frac{\sigma^2 r^2}{2} \cdot \left(1 + \frac{10V_f}{\gamma^2 \sigma^2} \frac{r}{1 - \frac{5}{\gamma} r}\right).$$

Denote $K := \frac{10V_f}{\gamma^2 \sigma^2}$, and $K' := K - \frac{5}{\gamma}$, then $K \geq \frac{5}{\gamma}$, $K' \geq 0$ (by Proposition 1.5), and

$$\begin{aligned} \beta_0(P(r)) &\leq 1 + \frac{\sigma^2 r^2}{2} \cdot \left(1 + K \frac{r}{1 - \frac{5}{\gamma} r}\right) = 1 + \frac{\sigma^2}{2} \cdot \frac{r^2 (1 + K' r)}{1 - \frac{5}{\gamma} r} \\ &\leq \exp\left(\frac{\sigma^2}{2} \cdot \frac{r^2 (1 + K' r)}{1 - \frac{5}{\gamma} r}\right). \end{aligned}$$

Finally, we apply (5.66), with the choice of r as the positive solution of

$$\frac{\sigma^2}{2} \cdot \frac{r^2 (1 + K' r)}{1 - \frac{5}{\gamma} r} = \frac{rt}{2}.$$

Solving this quadratic equation gives

$$r = -\frac{\sqrt{(\sigma^2 + \frac{5}{\gamma} t)^2 + 4\sigma^2 K' t} - (\sigma^2 + \frac{5}{\gamma} t)}{2\sigma^2 K'},$$

noticing that with this choice of r , $rt - \log(\beta_0(P(r))) \geq \frac{rt}{2}$, proves the result. \square

Proof of Theorem 2.8. The proof is based on the trick of [Janson \(2004\)](#), as in the proof of Corollary 2.4.

Denote $t_n = \sum_{i=1}^n f(X_i)$, then, by page 857 of [Lezaud \(1998a\)](#) and page 52 of [Lezaud \(1998b\)](#), in the case of initial distribution q , we have

$$\mathbb{E}_q \exp(rt_n) = q^T P(r)^n \mathbf{1} \leq N_q \| (P(r)^*)^n P(r)^n \|_{2 \rightarrow 2}^{1/2} \leq N_q \beta_0(r)^{n/2}, \quad (5.67)$$

where $\beta_0(r)$ denotes the largest eigenvalue of the operator $K(r) := P(r)^* P(r)$.

It is then shown that

$$\beta_0(r) \leq 1 + \frac{4V_f}{\gamma(K)} r^2 \left(1 - \frac{10r}{\gamma(K)} \right)^{-1}, \quad (5.68)$$

and the tail bound follows by Markov's inequality.

First, suppose that X_1, \dots, X_N is stationary, then $q = \pi$, and $N_q = 1$. Let $S = \sum_{i=1}^N f(X_i)$.

Let us fix some integer $k \geq 1$, and divide $f(X_1), \dots, f(X_N)$ into k parts:

$$(f(X_1), f(X_{k+1}), \dots), \dots, ((f(X_k), f(X_{2k}), \dots)).$$

Denote the sums of each part by S_1, \dots, S_k , then $S = \sum_{i=1}^k S_k$.

Suppose, without loss of generality, that N is divisible by k , then for each $i \leq k$, applying (5.67) and (5.68) gives

$$\mathbb{E} \exp(rS_i) \leq \exp \left(\frac{N}{k} \cdot \frac{2V_f r^2}{\gamma((P^*)^k P^k)} \left(1 - \frac{10r}{\gamma((P^*)^k P^k)} \right)^{-1} \right). \quad (5.69)$$

Jensen's inequality shows that $\mathbb{E} \exp(rS) \leq \frac{1}{k} \sum_{i=1}^k \mathbb{E} \exp(rkS_i)$, thus

$$\mathbb{E} \exp(rS) \leq \exp \left(N \cdot \frac{2V_f r^2}{\gamma((P^*)^k P^k)/k} \left(1 - \frac{10r}{\gamma((P^*)^k P^k)/k} \right)^{-1} \right). \quad (5.70)$$

This means that the statement of Theorem 3.3. of [Lezaud \(1998a\)](#) holds with $\gamma(K)$ replaced by $\gamma((P^*)^k P^k)/k$. Optimizing in k gives the result. The non-stationary case can be handled the same way as in Corollary 2.10. \square

Acknowledgements

The author thanks Doma Szász and Mogyi Tóth for infecting him with their enthusiasm of probability. He thanks his thesis supervisors, Louis Chen and Adrian Röllin, for the opportunity to study in Singapore, and their useful advices. Finally, many thanks to my brother, Roland Paulin, for the enlightening discussions.

References

- ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** no. 34, 1000–1034. . [MR2424985 \(2009j:60032\)](#)

- AJTAI, M., KOMLÓS, J. and SZEMERÉDI, E. (1983). An $O(n \log n)$ sorting network. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing* 1–9. ACM.
- AJTAI, M., KOMLÓS, J. and SZEMERÉDI, E. (1987). Deterministic simulation in LOGSPACE. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing* 132–140. ACM.
- BARBOUR, A. D. and CHEN, L. H. Y., eds. (2005). *An introduction to Stein's method. Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore* 4. Singapore University Press, Singapore. Lectures from the Meeting on Stein's Method and Applications: a Program in Honor of Charles Stein held at the National University of Singapore, Singapore, July 28–August 31, 2003. . [MR2235447 \(2007j:60001\)](#)
- BOLTHAUSEN, E. (1980). The Berry-Esseen theorem for functionals of discrete Markov chains. *Z. Wahrsch. Verw. Gebiete* **54** 59–73. . [MR595481 \(82a:60024\)](#)
- BOUCHERON, S., LUGOSI, G. and MASSART, P. (2009). On concentration of self-bounding functions. *Electron. J. Probab.* **14** no. 64, 1884–1899. [MR2540852 \(2010k:60058\)](#)
- BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. . [MR2123200 \(2006a:60024\)](#)
- BUBLEY, R., DYER, M. et al. (1997). Path coupling, Dobrushin uniqueness, and approximate counting. *Research Report Series - University of Leeds School of Computer Studies LU SCS RR*. Available at http://reference.kfupm.edu.sa/content/p/a/path_coupling__dobrushin_uniqueness__and_957
- CHATTERJEE, S. (2005). *Concentration inequalities with exchangeable pairs*. Thesis (Ph.D.)–Stanford University, Available at <http://arxiv.org/abs/math.PR/0507526>. [MR2707160](#)
- CHATTERJEE, S. (2007). Stein's method for concentration inequalities. *Probab. Theory Related Fields* **138** 305–321. . [MR2288072 \(2008e:60038\)](#)
- CHAWLA, S. (2010). CS880: Approximations Algorithms Lecture notes. Available at <http://pages.cs.wisc.edu/~shuchi/courses/880-S07/scribe-notes/lecture25.pdf>.
- CHAZOTTES, J.-R. and REDIG, F. (2009). Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.* **14** no. 40, 1162–1180. [MR2511280 \(2010g:60039\)](#)
- CHAZOTTES, J. R., COLLET, P., KÜLSKE, C. and REDIG, F. (2007). Concentration inequalities for random fields via coupling. *Probab. Theory Related Fields* **137** 201–225. . [MR2278456 \(2008i:60167\)](#)
- CHEN, L. H. Y., FANG, X. and SHAO, Q. M. (2009). From Stein identities to moderate deviations. *arXiv preprint arXiv:0911.5373*.
- CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q.-M. (2011). *Normal approximation by Stein's method. Probability and its Applications (New York)*. Springer, Heidelberg. [MR2732624](#)
- DEMBO, A. (1997). Information inequalities and concentration of measure. *Ann. Probab.* **25** 927–939. . [MR1434131 \(98e:60027\)](#)
- DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial methods in density estimation. Springer Series in Statistics*. Springer-Verlag, New York. . [MR1843146 \(2002h:62002\)](#)
- DIACONIS, P. (2009). The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc.*

- (N.S.) **46** 179–205. . [MR2476411 \(2010b:60204\)](#)
- DIACONIS, P., HOLMES, S. and MONTGOMERY, R. (2007). Dynamical bias in the coin toss. *SIAM Rev.* **49** 211–235. . [MR2327054 \(2009a:62016\)](#)
- DUBHASHI, D. P. and PANCONESI, A. (2009). *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge. . [MR2547432](#)
- DUDLEY, R. M. (2011). The Dvoretzky-Kiefer-Wolfowitz inequality with sharp constant: Massart’s 1990 proof. Seminar lecture, available at <http://math.mit.edu/~rmd/Research/semtalk.pdf>.
- DUDLEY, R. M. (2012). *Uniform Central Limit Theorems, 2nd edition*. Draft version available at <http://math.mit.edu/~rmd/998/>.
- GILLMAN, D. (1998). A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.* **27** 1203–1220. . [MR1621958 \(99e:60076\)](#)
- GYORI, B. and PAULIN, D. (2012). Non-asymptotic confidence intervals for MCMC. *arXiv preprint*.
- HOORY, S., LINIAL, N. and WIGDERSON, A. (2006). Expander graphs and their applications. *Bull. Amer. Math. Soc. (N.S.)* **43** 439–561 (electronic). . [MR2247919 \(2007h:68055\)](#)
- JANSON, S. (2004). Large deviations for sums of partly dependent random variables. *Random Structures Algorithms* **24** 234–248. . [MR2068873 \(2005e:60061\)](#)
- JERRUM, M. and SINCLAIR, A. (1996). The Markov chain Monte Carlo method: an approach to approximate counting and integration. *Approximation algorithms for NP-hard problems* 482–520.
- KAHALE, N. (1997). Large deviation bounds for Markov chains. *Combin. Probab. Comput.* **6** 465–474. . [MR1483429 \(98g:60120\)](#)
- KONTOROVICH, L. (2006). Measure concentration of hidden Markov processes. *arXiv preprint math/0608064*.
- KONTOROVICH, L. (2007). *Measure Concentration of Strongly Mixing Processes with Applications*. Ph.D. dissertation, Carnegie Mellon University, Available at <http://www.cs.bgu.ac.il/~karyeh/thesis.pdf>.
- KONTOROVICH, A. and WEISS, R. (2012). Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *arXiv preprint arXiv:1207.4678*.
- KOWALSKI, E. (2011). Documents for expander graphs course. Available at <http://www.math.ethz.ch/~kowalski/expanders.html>.
- KVAM, P. and SOKOL, J. S. (2006). A logistic regression/Markov chain model for NCAA basketball. *Naval research Logistics (NrL)* **53** 788–803.
- LEDOUX, M. (2001). *The concentration of measure phenomenon*. *Mathematical Surveys and Monographs* **89**. American Mathematical Society, Providence, RI. [MR1849347 \(2003k:28019\)](#)
- LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach spaces. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]* **23**. Springer-Verlag, Berlin. Isoperimetry and processes. [MR1102015 \(93c:60001\)](#)
- LEÓN, C. A. and PERRON, F. (2004). Optimal Hoeffding bounds for discrete reversible Markov chains. *Annals of Applied Probability* 958–970.
- LEVIN, D. A., PERES, Y. and WILMER, E. L. (2009). *Markov chains and mixing times*. American Mathematical Society, Providence, RI. With a chapter by James G. Propp and

- David B. Wilson. [MR2466937 \(2010c:60209\)](#)
- LEZAUD, P. (1998a). Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* **8** 849–867. . [MR1627795 \(99f:60061\)](#)
- LEZAUD, P. (1998b). *Etude quantitative des chaînes de Markov par perturbation de leur noyau*. Thèse doctorat mathématiques appliquées de l'Université Paul Sabatier de Toulouse, Available at http://pom.tls.cena.fr/papers/thesis/these_lezaud.pdf.
- LEZAUD, P. (2001). Chernoff and Berry-Esséen inequalities for Markov processes. *ESAIM Probab. Statist.* **5** 183–201. . [MR1875670 \(2003e:60168\)](#)
- LINDVALL, T. (1992). *Lectures on the coupling method*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication. [MR1180522 \(94c:60002\)](#)
- LOVÁSZ, L. and WINKLER, P. (1998). Mixing times. *Microsurveys in discrete probability* **41** 85–134.
- LUBETZKY, E. and SLY, A. (2009). Cutoff for the Ising model on the lattice. *Inventiones Mathematicae* 1–37.
- MANN, B. W. (1996). *Berry-Esseen central limit theorems for Markov chains*. ProQuest LLC, Ann Arbor, MI Thesis (Ph.D.)—Harvard University. [MR2694356](#)
- MARTON, K. (1986). A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory* **32** 445–446. . [MR838213 \(87e:94018\)](#)
- MARTON, K. (1996a). Bounding \bar{d} -distance by informational divergence: a method to prove measure concentration. *Ann. Probab.* **24** 857–866. . [MR1404531 \(97f:60064\)](#)
- MARTON, K. (1996b). A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **6** 556–571. . [MR1392329 \(97g:60082\)](#)
- MARTON, K. (1997). Erratum to: “A measure concentration inequality for contracting Markov chains” [*Geom. Funct. Anal.* **6** (1996), no. 3, 556–571; [MR1392329 \(97g:60082\)](#)]. *Geom. Funct. Anal.* **7** 609–613. . [MR1466340 \(98h:60096\)](#)
- MARTON, K. (1998a). Measure concentration for a class of random processes. *Probab. Theory Related Fields* **110** 427–439. . [MR1616492 \(99g:60074\)](#)
- MARTON, K. (1998b). On a measure concentration of Talagrand for dependent random variables. *Unpublished manuscript*.
- MARTON, K. (2003). Measure concentration and strong mixing. *Studia Sci. Math. Hungar.* **40** 95–113. . [MR2002993 \(2004f:60042\)](#)
- MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283. [MR1062069 \(91i:60052\)](#)
- MASSART, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–884. . [MR1782276 \(2001m:60038\)](#)
- MASSART, P. (2007). *Concentration inequalities and model selection*. *Lecture Notes in Mathematics* **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. [MR2319879 \(2010a:62008\)](#)
- MCDIARMID, C. (2002). Concentration for independent permutations. *Combin. Probab. Comput.* **11** 163–178. . [MR1888907 \(2002m:60016\)](#)
- MITZENMACHER, M. and UPFAL, E. (2005). *Probability and computing*. Cambridge University Press, Cambridge. Randomized algorithms and probabilistic analysis. [MR2144605 \(2006d:68002\)](#)

- SAMSON, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ -mixing processes. *Ann. Probab.* **28** 416–461. . [MR1756011 \(2001d:60015\)](#)
- TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.* 81 73–205. [MR1361756 \(97h:60016\)](#)
- TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. . [MR1419006 \(99b:60030\)](#)
- TALAGRAND, M. (2011). *Mean field models for spin glasses. Volume I. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]* **54**. Springer-Verlag, Berlin. Basic examples. . [MR2731561 \(2012c:82036\)](#)
- TAO, T. (2010). 254B, Notes 1: Basic theory of expander graphs. Available at <http://terrytao.wordpress.com/2011/12/02/245b-notes-1-basic-theory-of-expander-graphs/>
- VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.